Novel Bayesian inference in epidemics – model assessment and integrating epidemiological and genetic data

Siu Yin Lau

Submitted for the degree of

Doctor of Philosophy

Heriot-Watt University

Department of Actuarial Mathematics & Statistics,

School of Mathematical and Computer Sciences.

March, 2015

Abstract

Work in this thesis represents advances in addressing two key challenges in epidemiological and ecological modelling: the lack of an effective and easily deployable model-assessment tool and a statistically sound joint inferential framework for epidemic and evolutionary processes. Firstly, we present a novel statistical framework that combines classical and Bayesian reasoning in testing for mis-specifications of a spatio-temporal model by investigating the consistency of so-called latent residuals with a known sampling distribution using a classical hypothesis test. Second, we devise a statistically sound Bayesian framework which facilitates the integration of epidemiological and genetic data; specifically, we demonstrate how the transmitted sequences can be effectively imputed so that the transmission dynamics of the joint epidemic and evolution process can be accurately recovered and also any unsampled infected hosts can be naturally accommodated in the analysis. The new methodology we propose are assessed using simulation studies and they are applied to two real-world epidemic datasets which respectively describe the spread of an invasive plant and foot-and-mouth disease in the UK, which shows that they may greatly enhance our ability to understand the transmission dynamics of disease and therefore lead to more efficient disease management.

Acknowledgements

Many people have helped me get this thesis done in their unique ways. Firstly, I would like to thank my supervisors Gavin, Glenn and George. Despite their involvements in various projects and duties, they have been always available with useful advice. In particular, thank them for commenting and proofreading this thesis. I once considered that decision of pursuing a PhD in a remote country far from my hometown Hong Kong as a huge and slightly risky "investment". Now, after the three-year time spent with this fantastic group and with the beautiful city, Edinburgh, I find the decision wise and rewarding. I would also like to thank my officemate Jeff who has been helpful and patient in answering my questions concerning computer programming. Also thank others in the office for they have made the office a friendly place for work. Thank Iain for his help for I.T. issues.

A special thanks must be given to my wife Angeline for her love, support and understanding. Without these, it would not have been possible for me to finish my PhD in three years. Thank her for bearing our lovely daughter Isla who is now almost nine-month old. Also thank her for bearing with me when I need to work and she has to take care of the newborn by herself.

Contents

1 Introduction **1**

1.1 An overview of the challenges in epidemic modelling 1

1.1.1 Inferring transmission dynamics 1

1.1.2 Joint analysis of epidemiological and sequence data 3

1.2 An overview and outline of the thesis 5

2 Background **7**

2.1 Epidemic models 7

2.1.1 Mathematical representations and assumptions 8

2.1.2 Sellke thresholds and simulation of an epidemic 10

2.2 Molecular evolution of pathogens 12

2.2.1 Mutations, DNA and RNA 12

2.2.2 Markov process for nucleotide mutations 12

2.2.3 Common models 14

2.2.4 Complexity of molecular evolution 15

2.3 Bayesian inference 16

2.3.1 Why Bayesian? 16

2.3.2 Noninformative prior 17

2.3.3 Markov chain Monte Carlo (MCMC) 20

2.4 Partially observed epidemic process and data augmentation 23

2.5 Bayesian model selection techniques 25

2.5.1 Bayes factor 25

2.5.2 Deviance Information Criterion (DIC) 25

2.5.3 Posterior predictive checks 27

3 A novel model assessment framework for spatio-temporal models in epidemiology and ecology **29**

3.1 Introduction 29

3.2 Model and Methods 31

3.2.1 Spatio-temporal stochastic model 31

3.2.2 Latent residuals 31

3.3 Reconstructing the epidemic using the residual process 34
3.3.1 Transition probabilities 34
3.3.2 Sellke thresholds, construction of the exposure times and the infection links 35
3.3.3 Construction of the sojourn time 36
3.3.4 Detailed algorithm for simulating epidemics utilizing the residual process 37
3.3.5 Bayesian inference and model assessment 38
3.3.6 Interpretation of latent-residual tests 40
3.3.7 Likelihood 41
3.3.8 Estimation 42
3.4 Simulated example 44
3.4.1 Rationale for ordering the infection links 49
3.4.2 Diagnosing model mis-specification 51
3.4.3 Comparison with common Bayesian model checking techniques 51
3.5 Posterior predictive checking based on spatial autocorrelation analysis 53
3.6 Case Study: Spread of Giant Hogweed in Great Britain 55
3.6.1 Model assessment and implications for control strategies . . . 59
3.6.2 Comparison with DIC and posterior predictive checks 61
3.6.3 Details of model formulation and posterior distributions of model parameters 62
3.7 Limitations and extensions 64
3.7.1 A confounding issue 64
3.7.2 Sequential latent-residual testing 68
3.8 Discussion 72

4 A new Bayesian computational method for the integrated analysis of epidemic and genetic data 74
4.1 Introduction 74
4.2 Model and methods 77
4.2.1 The stochastic epidemic process 77
4.2.2 The stochastic evolutionary process 78
4.2.3 Modelling background pathogen and multiple clusters 78
4.2.4 Likelihood 79
4.2.5 Bayesian inference and MCMC 83
4.2.6 Other details of the MCMC algorithm 89
4.3 Simulation studies 92
4.3.1 Inference for epidemics with multiple clusters 92
4.3.2 Inference for epidemics with single cluster 110

4.4 More simulated epidemics 113
4.5 Contribution of genetic data to model assessment 119
4.6 Case study: spread of foot-and-mouth disease virus in UK (Darlington, Durham county, 2001) 120
4.6.1 Revisiting the inference of the transmission dynamics 121
4.6.2 Inclusion of unreported susceptibles 124
4.7 Validation of the methodology 127
4.7.1 Fitting a full model to epidemic data 127
4.7.2 Posterior distribution of parameter p for the FMD outbreak (Darlington, 2001) 127
4.8 Discussion 129

5 Conclusion and future work **132**
5.1 Conclusion to the thesis 132
5.2 Future developments for the residual testing 133
5.2.1 A sequential approach 133
5.2.2 Exposure Time Residuals (ETR) 134
5.3 Future developments for the joint analysis of epidemic and genetic data 135
5.3.1 Modelling within-host dynamics 135
5.3.2 Alternative sampling schemes of genetic data 137

Appendix A A R package for latent residuals test **138**
A.1 A brief description 138
A.2 Flexibility of the package 138

List of Figures

1.1 An epidemic simulated from an exponentially-bounded spatial kernel. 3
1.2 A phylogenetic tree and a transmission event. 5

2.1 Distribution of a random variable obtained from transforming a uniform prior. 18

3.1 A graphical comparison between the centered and the non-centered parameterisation. (a) The centered parameterisation; (b) The non-centered parameterisation/ functional model representation. 32
3.2 An illustration of (a subset of) the observed data y in the simulation (replicate 1) in the 2000×2000 square area in the form of a sequence 'snapshots' of the system state at particular times. 45
3.3 Posterior distributions of model parameters for models fitted to the simulated data. 46
3.4 Traceplots of the posterior samples of model parameters (with burn-in length 10,000) obtained from fitting the correct model to the simulated data. 47
3.5 A schematic representation of the relative strength of interaction of these kernels. 50
3.6 The distributions of a subset of imputed $\widetilde{r}_2$ whose $P(\widetilde{r}_2) < 0.05$ under two scenarios. 51
3.7 Posterior predictive distributions of Moran's I and Geary's c indexes obtained by simulating 1,000 epidemics from Model I and Model II respectively at time points $T_1 = 25$, $T_2 = 35$ and $T_3 = 45$. 53
3.8 Estimated spatial kernels from fitting Model I and Model II with kernel parameters set to posterior means. 54
3.9 Snapshots of the spread of giant hogweed in Great Britain taken at three distinct times: (a) 1970, (b) 1987 and (c) 2000. 56

3.10 Posterior distributions of the p-values from testing the sets of posterior samples of Infection Link Residual (ILR) imputed from MCMC chains (1,500 samples in each case) when fitting SI models, representing heterogeneous suitability, to the giant hogweed data with kernel A (model M1) and kernel B (model M2) respectively. 57
3.11 Estimated spatial kernels from fitting M1 and M2 and M3 to the giant hogweed data with kernel parameters set to posterior means. 60
3.12 Distributions of subsets of imputed ILR which lead to p-values less than 0.05 from M2. 60
3.13 The partition of Great Britain according to intersection with 65 concentric annuli 62
3.14 Distribution of the number of colonised sites within each ring region at the final observation time as predicted by models and the observed data. 63
3.15 Posterior distributions of transmission rates from a colonised site to a susceptible site from fitting M1 and M2 to the giant hogweed data . . 63
3.16 Posterior distributions of model parameters for models fitted to giant hogweed data where suitability of sites are considered. 65
3.17 Traceplots of the posterior samples of model parameters obtained from fitting model M1 to giant hogweed data where suitability of sites are considered (with burn-in length 10,000). 66
3.18 Traceplots of the posterior samples of model parameters obtained from fitting model M2 to giant hogweed data where suitability of sites are considered (with burn-in length 10,000). 67
3.19 Posterior distributions of suitability parameters in the model (with kernel A) fitted to the giant hogweed data in which sites are classified into three classes 67
3.20 Posterior distributions of the p-values from testing the sets of posterior samples of Infection Link Residual (ILR) imputed from MCMC chains in fitting different combinations of the spatial kernel and the latent period. 68
3.21 Posterior distributions of the p-values from testing the sets of posterior samples of Latent Time Residual (LTR) imputed from MCMC chains in fitting different combinations of the spatial kernel and the latent period. 69

4.1 The event of individual i infecting individuals j and k and the sampling of sequences on these individuals. 80
4.2 Illustration of the selection t_p (and the corresponding past sequence G_p) and t_f (and the corresponding past sequence G_f). 86

4.3 Posterior distributions of the overall coverage rate for the two multiple-cluster epidemics. (a) 3-cluster; (b) 6-cluster 94
4.4 Posterior distributions of the model parameters (with the *three-cluster* epidemic). 95
4.5 Posterior distributions of the model parameters (with the *six-cluster* epidemic). 96
4.6 Traceplots of the posterior samples of model parameters in the case with 100% sampling (with the *three-cluster* epidemic). Dotted lines represent the true values of the model parameters. 97
4.7 Traceplots of the posterior samples of model parameters in the case with 50% sampling (with the *three-cluster* epidemic). Dotted lines represent the true values of the model parameters. 98
4.8 Traceplots of the posterior samples of model parameters in the case with 100% sampling (with the *six-cluster* epidemic). Dotted lines represent the true values of the model parameters. 99
4.9 Traceplots of the posterior samples of model parameters in the case with 50% sampling (with the *six-cluster* epidemic). Dotted lines represent the true values of the model parameters. 100
4.10 Posterior distributions of model parameters and the cover rate from fitting the three-cluster epidemic data with sampling proportion 20% (assuming the latent period distribution is known). 101
4.11 Posterior distributions of model parameters and the cover rate from fitting the six-cluster epidemic data with sampling proportion 20% (assuming the latent period distribution is known). 102
4.12 Posterior *individual coverage* of the source of infection (see main text) in scenarios with sampling 100%, 80%, 50% and 0%. 103
4.13 Posterior *cluster identification rate* of the infections, within each actual cluster of the *six-cluster* epidemic. 105
4.14 Posterior (primary) *ancestor identification rate* of the infections, within each actual cluster of the *six-cluster* epidemic. 106
4.15 Posterior distributions of the model parameters for the epidemic with lower mutation rates. 107
4.16 Posterior distributions of the overall coverage rate for the epidemic with lower mutation rates. 108
4.17 Posterior individual coverage of the sources of infection for the epidemic with lower mutation rates in scenarios with sampling 100%, 50%, 10% and 0%. 108

4.18 Posterior cluster identification rate of the infections, within each actual cluster of the epidemic with lower mutation rates, in scenarios with sampling 100%, 50%, 10% and 0%. 109
4.19 Posterior (primary) ancestor identification rate of the infections, within each actual cluster of the epidemic with lower mutation rates, in scenarios with sampling 100%, 50%, 10% and 0%. 110
4.20 Posterior distributions of model parameters using the pseudo-likelihood approach. 111
4.21 Posterior distributions of the mutation rates (with the single-cluster epidemic). (a) $n = 1000$; (b) $n = 8000$ 112
4.22 Posterior distributions of the overall coverage rate (with the single-cluster epidemic). (a) $n = 1000$; (b) $n = 8000$ 113
4.23 Violin plots showing the posterior distributions of the model parameters (with the single-cluster epidemic and number of bases $n = 1000$). 114
4.24 Posterior distributions of the model parameters (with the single-cluster epidemic and number of bases $n = 8000$) 115
4.25 Posterior distributions of the mean latent period, denoted as μ_{lat}, and of the transition rate μ_1 and transversion rate μ_2. 122
4.26 Traceplots for the posterior samples of the mean latent period and of the transition rate and transversion rate. 122
4.27 (a) The transmission graph with highest posterior probability, 0.89; (b) The transmission graph with the second highest posterior probability, 0.08. 123
4.28 Posterior distributions of the full set of model parameters obtained from fitting the model to 10 independently simulated datasets obtained by adding 300 randomly assigned susceptible premises to the 2001 FMD data. 124
4.29 Posterior distributions of the full set of model parameters obtained from fitting the model to 10 independently simulated datasets obtained by adding 100 randomly assigned susceptible premises to the 2001 FMD data. 125
4.30 Posterior distributions of the full set of model parameters obtained from fitting the model to 10 independently simulated datasets obtained by adding 500 randomly assigned susceptible premises to the 2001 FMD data. 126
4.31 Comparisons between the posterior distributions of the coverage rates and of κ obtained from fitting two models, the full model (Scenario I) and the epidemic model (Scenario II), to the epidemic data (no sampled sequences). 128

5.1 An illustration of a Yule process. Starting from one strain, after going through 3 birth events at 3 nodes (indicated by the black dots), evolves to 4 strains (*a, b, c, d*) at time t. 135

Chapter 1

Introduction

1.1 An overview of the challenges in epidemic modelling

Mathematical and statistical modelling in epidemiological and ecological studies has evolved rapidly over the past decades, largely due to improved surveillance, which has led to much richer data, and the advancement of simulation-based statistical methods and computer power. Despite substantial developments, two key challenges are being presented to epidemic modellers. First, while there is a rich set of well-studied models, effective model assessment techniques are lacking. On the other hand, the rapidly growing availability and volume of data present both opportunities and challenges. In particular, sequence data of pathogens are becoming increasingly available due to the reduced cost of genome sequencing. These sequence data, embedding information of the evolutionary history of pathogens among the infected hosts in a population, aid to reveal greater detail on key aspects of the transmission dynamics of epidemics such as the transmission network than standard epidemiological data. A key challenge is to integrate standard epidemic data with these sequence data within a statistically sound framework.

1.1.1 Inferring transmission dynamics

Stochastic spatio-temporal models are playing an increasingly important role in epidemiological and ecological studies relating to transmission of diseases (Ster et al, 2009), invasion of alien species (Cook et al, 2007a) and population movements in response to climate changes (Walters et al, 2006) (see 2.1 in Chapter 2 for discussion of common spatio-temporal epidemic models). It is well known that the predicted

dynamics of such systems can be extremely sensitive to the choice of model, with consequent implications for the design of control strategies (Ster et al, 2009; Ferguson et al, 2001), but as yet there is a lack of effective model assessment tools described in the literature. For example, studies of foot and mouth disease have cited the importance of selecting between a long-tailed *spatial kernel* (see later) versus a localised spatial kernel (Keeling et al, 2003; Ferguson et al, 2001). Further model-choice problems arise in relation to the parametric form of the distributions of incubation and infectious periods in models of measles (Ferguson et al, 1997; Bolker and Grenfell, 1995), and in relation to diseases such as smallpox (Streftaris and Gibson, 2004b) and AIDS (Muñoz et al, 1997).

A major and unresolved issue in (spatio-temporal) epidemic model specifications is to choose an appropriate spatial kernel which describes the dependence of the infectious challenge from an infective to a susceptible site (detailed mathematical formulations are given in Chapter 3). In particular, there is a lack of tools to distinguish between a long-tailed and a short-tailed spatial kernel which can imply very distinct control strategies (see later). A short-tailed spatial kernel such as an exponentially-bounded kernel mostly allows short-distance interactions between infectives and susceptibles and hence result in a more localised and "continuous" spreading pattern. In contrast, it is well-known that a long-tailed spatial kernel such as a Cauchy-type kernel allows more frequent long-distance interactions and hence result in a more patchy pattern of spread, with apparently multiple foci. In fact, at sufficiently large scale either type of spatial kernel can lead to an emergent pattern looking like an epidemic emanating from a single focus (Shaw, 1995).

The spatial kernel has important implications for devising effective control strategies (Gibson, 1997; Keeling et al, 2003; Ferguson et al, 2001). For instance, in animal disease outbreaks (e.g., foot-and-mouth disease, a major animal disease which has caused significant economic losses in the UK), culling (i.e., the pre-emptive slaughter of all susceptible animals on farms around infection sites) and vaccination are two major measures aiming at reducing the density of susceptibles and hence limiting the chance of further spread of the disease in the population (Keeling et al, 2003; Ferguson et al, 2001). Optimal implementation of such strategies (e.g., how large a geographical area around infection sites should receive culling) clearly depends on, for example, how likely long-distance transmission is to occur, which is being implied by the spatial kernel.

Despite the very different underlying mechanisms they represent, distinguishing between different kernels is usually impeded by the complexity of epidemic dynamics. For example, a patchy pattern may also be observed by having a short-tailed kernel in combination with background infections/colonisations (see Figure 1.1). As a result,

approaches relying on analysing the spatial point patterns (Getis, 1991) may not be effective for distinguishing between models as two types of spatial kernel may be capable of producing similar patterns of spread (Gibson, 1997). *Posterior predictive checking* which determines goodness-of-fit by benchmarking summary statistics from the data against their posterior predictive distributions may not be sensitive as they merely reflect the averaging behaviour of competing models. Other approaches such as *Bayes factors* and *Deviance Information Criteria (DIC)* suffer from different key limitations. In particular, none of the existing approaches are designed to test the goodness-of-fit of the spatial kernel (or any particular components) in a general spatio-temporal epidemic model; rather, they provide an indication of the overall goodness-of-fit of the full model and are usually difficult to interpret. Illustrations of the limitations of these approaches are given in Chapter 2 and Chapter 3. Developing effective and more model-component targeted approaches is certainly warranted.

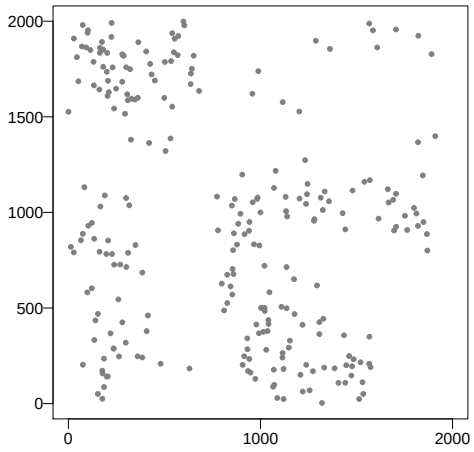


Figure 1.1: An epidemic simulated from a spatio-temporal model (see later chapters for more details) within a 2000×2000 square region using an exponentially-bounded spatial kernel and a relatively high background infection rate. Solid dots represent infected sites.

1.1.2 Joint analysis of epidemiological and sequence data

Historically, epidemiological data collected during epidemic outbreaks provide the main source of information from which to infer disease transmission dynamics. These data are based on clinical observation and diagnostic test results and typically include *observed* times of symptom onset or times of culling/removal.

In order to capture the transmission dynamics more adequately, we need to infer some key *unobserved* aspects of an epidemic model such as exposure times. Being the first key papers addressing this issue, Gibson and Renshaw (1998) and O'Neill and

Roberts (1999) demonstrated how unobserved components in a general $Susceptible-Exposed-Infectious-Recovered$ model (see also details of epidemiological models in Chapter 2) can be imputed in a Bayesian framework using a computationally-intensive numerical method called Markov chain Monte Carlo (MCMC). Considerable progress has also been made in developing subsequent statistical methods for inference from these partially observed epidemics (e.g., (Streftaris and Gibson, 2012)). In spite of the progress made, during a typical outbreak the epidemiological data available do not allow very precise inference of detailed aspects of disease transmission dynamics, for they only indirectly reflect the underlying contact structures, exposure times, and other information needed, for example, to infer the transmission network.

Another valuable source of data is genetic data of pathogens sampled from infected hosts during epidemic outbreaks. These data indicate the evolutionary history of pathogens and therefore carry information on relatedness of different infection events. Such data have become increasingly available in recent years (Rambaut et al, 2008; Cottam et al, 2008), which presents both opportunities and challenges to modellers. While they have great potential to be used to reveal more detailed aspects of the transmission dynamics, substantial developments of new statistical methodology are required to integrate them with traditional epidemic data in a statistically sound and hopefully a more powerful framework.

Various approaches have been proposed. Approaches that rely on reconstructing *phylogenetic trees* were first considered in several scenarios focusing on estimating the evolutionary aspects (e.g., the sequence ages) of pathogens (Shapiro et al, 2011; Rambaut et al, 2008). A phylogenetic tree (Figure 1.2 (a)) is a diagram that depicts the evolutionary relationships among species or entities (in our case we are interested in the relationship between pathogens sampled from different hosts which could be individuals or premises). These approaches may be inappropriate to infer the transmission dynamics, such as the transmission network, when sampled sequences include donor-recipient pairs. For example, let us suppose that a site A infected another site B. A sample from A was collected before the infection event and a sample from B was collected after the infection event (as shown in Figure 1.2 (b)), it is clear that these two samples form a donor-recipient pair, and it is inappropriate to assign them to the *tips* (also known as *leaves*) − instead, the sample from A is more appropriately to be assigned as an ancestor (e.g., along the branch or on the node) that descends to sample from B. A more recently developed approach is to consider the transmission network explicitly (Ypma et al, 2012; Morelli et al, 2012; Ypma et al, 2013; Jombart et al, 2014). However, these approaches do not account for *unobserved transmitted sequences* (solid circles coloured grey in Figure 1.2 (b)) between donor-recipient pairs which are unequivocally required to represent the transmission dynamics adequately

and they mostly rely on pseudo-likelihood approaches that cannot describe adequately the dependence between collected samples (solid circles coloured black in Figure 1.2 (b)). More details of the pseudo-likelihood approaches and other existing approaches are also discussed in Chapter 4. Further research in developing a more accurate approach for the joint inference of epidemic and genetic data is needed.

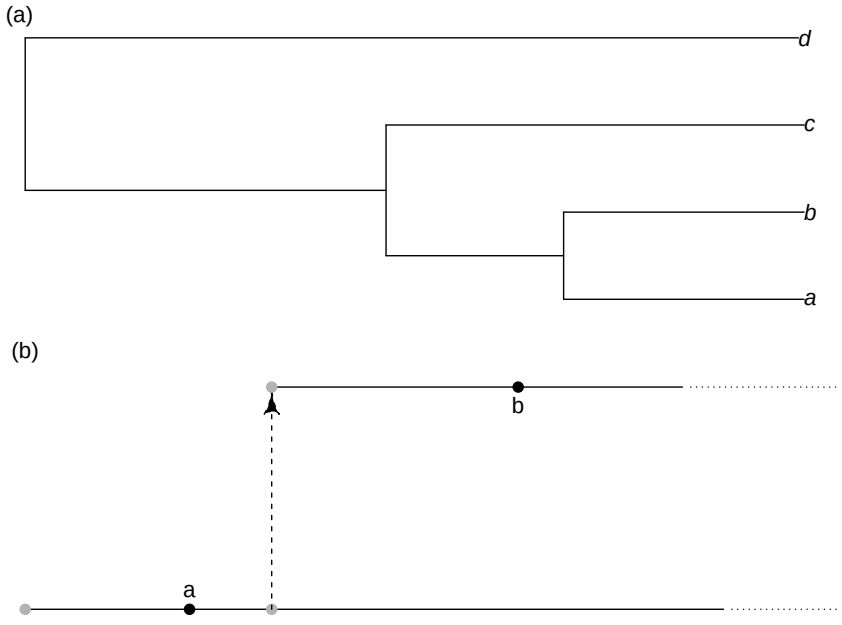


Figure 1.2: (a) A hypothetical phylogenetic tree constructed from samples a, b, c, d collected from sites A, B, C, D respectively: the tips or leaves of a phylogenetic tree represent the sampled pathogens and each node represents a common ancestor of the tips descended from it. (b) The event of site A infecting B and the sampling of sequences on these individuals. The sequence samples (coloured black) are observed and the transmitted sequences (coloured grey) are typically unobserved. Possible events on dotted lines are not shown.

1.2 An overview and outline of the thesis

This thesis aims to contribute to address these the challenges presented in section 1.1. Specifically, we first propose a novel Bayesian model assessment tool which is sensitive to mis-specifications of particular components of a general spatio-temporal model and is able to provide a statistically interpretable diagnosis of model specification. This framework requires and presents an innovative re-parametrisation scheme of epidemic processes involving a known residual processes independent of model assumptions. These residuals are then imputed in a Bayesian framework using MCMC. On applying a classical test for consistency with the known distribution of the imputed residuals a posterior distribution of *p-values* is generated, from which evidence against the

modelling assumptions can be discerned. The approach has its roots in the *posterior predictive p-value* proposed in Meng (1994), and extended in Streftaris and Gibson (2004a) and Gibson et al (2006). The key innovation in this work is to design the residual processes so that the resulting tests are sensitive to mis-specification of specific aspects of the model under consideration. In contrast to conventional posterior predictive checks relying on summary statistics of observed data, our approach utilises the posterior samples of the residual processes with a known distribution with which the test may be more sensitive to model mis-specifications.

Another key aim of the thesis is to devise a statistically sound method for integrating epidemiological and genetic data. As accounting for unobserved transmitted sequences requires a very high-dimensional model space and this imposes great challenges to statistical inference, alternative approaches, compromising the accuracy of inference, have been proposed. In this thesis, we show that it is feasible to impute the unobserved transmitted sequences between infected hosts, which has been the main difficulty in the joint inference of epidemic and genetic data. We also investigate comprehensively the values of genetic data in inferring the transmission dynamics. In particular, we demonstrate how genetic data may aid the estimation of epidemiological parameters, which has been largely ignored in the literature. We also demonstrate the practicality in using partial genomic data, which may bear important implications for future study designs.

The thesis is structured as follows. Chapter 2 gives an overview of current important statistical models and methods in epidemic modelling which are to be used in the succeeding chapters. The new methodology we propose is explained in detail in Chapter 3 and Chapter 4 and is assessed using simulation studies. Two real-world epidemic datasets which respectively describes the spread of an invasive plant and foot-and-mouth disease in the UK are analysed in these two chapters demonstrating the developed methodology. We discuss potential future developments of this work in Chapter 5.

Chapter 2

Background

Chapter summary

In this chapter we give an overview of existing epidemic and evolutionary models and some methodology for key statistical inference which will be used in later chapters. We also highlight the limitations of current model selection techniques before we proceed to the next chapter where a novel model assessment framework is proposed.

2.1 Epidemic models

Stochastic spatio-temporal models are playing an increasingly important role in epidemiological and ecological studies relating to transmission of diseases (Ster et al, 2009; Ferguson et al, 2001), invasion of alien species (Lau et al, 2014b; Cook et al, 2007a) and population movements in response to climate changes (Walters et al, 2006). In this section we discuss some of the key model structures and the related underlying assumptions.

A standard approach in epidemiological modelling is to describe disease progression in a sequence of compartments, which are usually consistent with clinical evidence. A typical example is the $SEIR$ model with susceptible (S), exposed (E), infectious (I) and removed/recovered (R) compartments. The transition of a susceptible individual to class E is determined by an assumed infection process (see later) in which infectious individuals play a key role. Then an exposed individual stays in class E before becoming infectious for a duration called *latent period*. An infectious individual then stays infectious for an *infectious period* before it is removed/recovered from the population. The notion of having sojourn times in class E and class I are supported by clinical evidence. For example, an individual exposed to influenza virus may not

be infectious for a period of time due to the low viral shedding at the beginning of infection and only remain infectious for a length of time before recovery (i.e., removal) (Hall et al, 1979). The $SEIR$ model is also commonly used to describe disease progression in animals in which the symptom onset times (e.g., the lesion times in foot-and-mouth disease) are often taken to be estimates of the time exposed individuals become infectious (with generally unknown exposure times) (Keeling et al, 2003; Ferguson et al, 2001). In animal disease epidemics, removal times usually correspond to culling times. For plant diseases, SIR and SI models are instead more common due to the aggressive nature of these diseases (Lau et al, 2014b; Cook et al, 2007a). Other settings are possible. For example, SIS model allows reinfections to occur when levels of infection are high or vaccination fails to protect (Gomes et al, 2004) − an important example is the dengue fever where an individual recovered from one strain may be infected by another strain (Kawaguchi et al, 2003). Infections may also be classified into asymptomatic or symptomatic which may have different infectiousness, which can be modelled by having an exposed individual entering two infectious classes (Mathews et al, 2007).

2.1.1 Mathematical representations and assumptions

Consider an $SEIR$ model and let $S(t)$, $E(t)$, $I(t)$ and $R(t)$ be the respective numbers in these classes at time t. The progression of individuals through compartments is often modelled as a Markov process with the occurrence of future events being *independent* of past events given the current state of the system. Examples of using Markovian models include measles and tuberculosis (Cauchemez and Ferguson, 2008; Debanne et al, 2000). A common mathematical formulation for probabilities of the three possible transition events to occur within time interval $[t, t + dt)$ is as follows:

$$\begin{aligned} &Pr(S(t+dt) = S(t) - 1, E(t+dt) = E(t) + 1) = \beta I(t)S(t)dt + o(dt) \\ &Pr(E(t+dt) = E(t) - 1, I(t+dt) = I(t) + 1) = \lambda E(t)dt + o(dt) \\ &Pr(I(t+dt) = I(t) - 1, R(t+dt) = R(t) + 1) = \sigma I(t)dt + o(dt), \end{aligned} \tag{2.1}$$

with β being the contact rate between an infectious and an exposed individual and λ and σ being the transition rates for the corresponding transitions to class I and class R.

The usage of constant λ and σ equivalently assumes exponentially distributed sojourn times in respective classes. While this assumption is mathematically convenient, it is

not necessarily biologically realistic. A less restrictive, and perhaps more realistic, assumption is to allow sojourn time distributions (e.g., a Gamma distribution) that can exhibit a modal value of time. Two common alternatives to exponentially distributed sojourn times are Weibull and Gamma distributions.

The assumption of *homogeneous mixing* among individuals in the population, implied by the first line of Equation 2.1 where each infectious individual exerts an equal infectious challenge to any given susceptible, may be inappropriate particularly in a spatial setting. For example, for a spatially stratified population where individuals may be divided into classes, a simple and more realistic extension is to assume individuals mix homogeneously within a particular class, but mix with a lesser extent with individuals in other classes (Morris, 1996; Watson, 1972). This introduces $n \times (n-1)$ levels of β, where n is the number of classes.

Instead of introducing additional levels of β, another natural way to account for heterogeneity in contact is to use a *spatial kernel* function to characterise the dependence of the infectious challenge from infective i to susceptible j as a function of distance d_{ij} between i and j. Using a spatial kernel is natural for spatially distributed and interacted populations. For example, the monotonically decreasing property of a spatial kernel function (e.g., an exponentially-bounded or power-law kernel function) renders it a natural candidate to describe the interactions/transmissions between an infectious premises to a susceptible premises (e.g., premises can be farms in outbreaks of foot-and-mouth disease or trees in outbreaks of Citrus greening disease (Ster et al, 2009; Keeling et al, 2003; Parry et al, 2014)). In Chapter 3 and Chapter 4 we consider a general spatio-temporal SEIR model (see below) in which the spatial connectivity is modelled by the spatial kernel.

Spatio-temporal SEIR model We consider a broad class of spatio-temporal stochastic models exemplified by the SEIR epidemic model with susceptible (S), exposed (E), infectious (I) and removed (R) compartments. Suppose that we have a spatially distributed population indexed 1, 2, ..., denote $\xi_S(t)$, $\xi_E(t)$, $\xi_I(t)$ and $\xi_R(t)$ as the set of indices for individuals who are in class S, E, I and class R respectively at time t, and let $S(t)$, $E(t)$, $I(t)$ and $R(t)$ be the respective numbers of individuals in these classes at time t. Then individual $j \in \xi_S(t)$ becomes exposed during $[t, t+dt)$ with probability

$$p(j,t) = \{\alpha + \beta \sum_{i \in \xi_I(t)} K(d_{ij}, \kappa)\}dt + o(dt), \tag{2.2}$$

where α represents a primary infection rate and β is a contact parameter. The term $K(d_{ij}, \kappa)$ (parametrized by the kernel parameter κ) characterises the dependence of the infectious challenge from infective i to j as a function of distance d_{ij} and is

known as the *spatial kernel function*. Following exposure, the random times spent by individuals in classes E and I are modelled using a suitable distribution such as a Gamma or a Weibull distribution (Streftaris and Gibson, 2004b; Valleron et al, 2001; Anderson, 1988). Specifically, we use a $Gamma(\mu, \sigma^2)$ parameterised by the mean, μ, and variance, σ^2, for the random time x spent in class E with density function

$$f_E(x; \mu, \sigma^2) = \frac{1}{(\sigma^2/\mu)^{\mu^2/\sigma^2}\Gamma(\mu^2/\sigma^2)} x^{\mu^2/\sigma^2 - 1} e^{-x\mu/\sigma^2}. \tag{2.3}$$

For the random time x spent in class I we use a $Weibull(\gamma, \eta)$ parameterised by the shape and scale with density function

$$f_I(x; \gamma, \eta) = (\eta/\gamma)(x/\gamma)^{\eta-1} e^{-(x/\gamma)^\eta}. \tag{2.4}$$

All sojourn times are assumed independent of each other given the model parameters.

As we have discussed in Chapter 1, it can be difficult to specify an adequate spatial kernel and different spatial kernels can correspond to very distinct assumptions over the transmission mechanism. Given the flexibility to choose kernel parameters when fitting models to data two rather different kernels can give apparently decent representations of the data, but these they may in fact lead to very different predictions of the subsequent behaviour of the epidemic (this issue is further discussed in 2.5.3 and demonstrated in Chapter 3). We aim to address this issue of model choices in Chapter 3.

2.1.2 Sellke thresholds and simulation of an epidemic

Sellke (1983) demonstrated the equivalence of a threshold model and a standard time-homogeneous Markov process (e.g., Equation 2.1) if the thresholds are assumed to be independent and identically distributed exponential random variables (with mean equal to 1). The key assumption in this threshold model is that an individual possesses a *random resistance to infection* which is termed as the *Sellke threshold*. That is, the individual would only be infected when the *infective pressure* (see Equation 2.5) from all infectious individuals reaches the Sellke threshold assigned. Sellke's construction can be utilised to simulate an epidemic from an SEIR model.

Algorithm

The core idea is that, given the history of the population to date, we simulate the next event time for each individual and locate the *earliest* event among these simulated events as the next event of whole epidemic process. The algorithm is explained in detail in the following paragraphs.

Suppose we have a spatially distributed population with defined locations of the individuals. For those uninfected, we calculate the additional *infectious pressure* required for each of them to get infected at the time that the current event happens. Denote $c_j(t)$ and q'_j to be respectively the additional infectious pressure required (at time t) and the threshold of individual j. We have

$$c_j(t) = q'_j - \alpha t - \beta \sum_{i \in \mathbf{C}} K(d_{ij}; \kappa) \Delta_{ij} \tag{2.5}$$

where t is the current event time, $\mathbf{C}$ is the collection of indices of those having been infectious that may or may not have recovered. If the exponential threshold is to be used, the set $\mathbf{C}$ might only contain those remaining infectious at time t. Δ_{ij} represents the length of time for which individual j is exposed to individual i in set $\mathbf{C}$. Δ_{ij} would be taken as the $t - s_i$ when individual i has not recovered by time t or $r_i - s_i$ when it has recovered. Here s_i and r_i are the times of becoming infectious and recovered for individual i respectively. After calculating the required additional infective pressure for each susceptible to get infected, the calculation of each of their next event times (the infection times), E_j, would be as following,

$$E_j = \frac{c_j(t)}{\alpha + \beta \sum_{i \in \mathbf{D}} K(d_{ij}; \kappa)} + t \tag{2.6}$$

where $\mathbf{D}$ is the set of indices of those remaining infectious at time t. Note that we are assuming no other changes happen before E_j. For those in class E or I, the times of the next event would be determined by the respective waiting time distribution (e.g., a Gamma or a Weibull distribution).

The set of susceptibles and the sets $\mathbf{C}$ and $\mathbf{D}$ would be updated after each simulated event. The simulation stops either when all individuals have been infected/recovered, or the time of the next event exceeds a pre-assigned maximum value of time, or when other pre-assigned criteria are met.

2.2 Molecular evolution of pathogens

Knowledge of molecular evolution greatly augments the understanding of epidemic outbreaks, particularly for epidemics with fast-evolving viruses. For example, the genome data sequenced from farms have been shown to be invaluable for inferring the transmission network of foot-and-mouth disease outbreaks (Morelli et al, 2012; Ypma et al, 2013) which has been difficult to estimate with epidemic data alone. Essential concepts and models for developing the work in Chapter 4 are given in this section. The reader is invited to refer to Salemi and Vandamme (2003); Yang (2006) for a more detailed introduction.

2.2.1 Mutations, DNA and RNA

The genome, where the genetic information of pathogens is encoded, is mostly in the form of DNA (deoxyribonucleic acid) and sometimes in RNA (ribonucleic acid) (e.g., bacterial DNA versus RNA viruses). DNA is a double helix which consists of two strands which carry genetic information. On one side of each strand there is a chain of *nucleotide bases*. On the other side of the strand (i.e., the backbone), there are *deoxyribose moieties* which are joined by phosphodiester linkages. There are four types of nucleotide bases, namely *adenine (A), guanine (G), thymine (T) and cytosine (C)*. The first two types of nucleotide bases are grouped under the category *purines* and the last two under *pyrimidines*. RNA is very similar to DNA molecule but it is single-stranded, with thymine (T) in DNA being replaced by *uracil (U)*.

Nucleotide mutations refer to the alteration of a nucleotide base by chemical reactions from mutagens (e.g., pollutants in the environment and UV light) or spontaneous mutation during the genetic information duplication. Point mutations which do not result in amino-acid [1] changes are called *synonymous mutations (silent mutations)* and those that do are called *nonsynonymous mutations*. Mutations to nucleotide bases within the same category are called *t*ransitions. Mutations between two categories are called *transversions*. Transversions usually have larger impact as they result in higher degrees of structural change.

2.2.2 Markov process for nucleotide mutations

The evolutionary process of the pathogen is often modelled at the level of nucleotide substitutions. Here *mutations* and *substitutions* are used interchangeably but we

[1] Simply speaking, each triplet of nucleotide bases (e.g., AUG) represents an amino acid.

note that this practice may not be appropriate in some contexts. For example, in a discussion involving *fixation events*, which refers to the scenarios that the change of genome resulting from particular mutations becomes fixed in the population, they have markedly different meanings as not all mutations will lead to fixations (i.e., substitutions).

Taking RNA viruses as an example, a continuous-time reversible Markov process taking state values (i.e., nucleotide bases) from the set $\omega_N = \{A, C, G, U\}$ can be defined as follows. Let μ_{xy} be the transition rate between any two states $x \in \omega_N$ and $y \in \omega_N$, and also let $\mathbf{F}(t) = (f_A(t), f_C(t), f_G(t), f_U(t))^t$ be the probabilities for a particular nucleotide position at state A, C, G and U respectively at time t.

A Markovian evolutionary process can be described by the equations

$$f_A(t+dt) = f_A(t) - \mu_A f_A(t) + \sum_{x \neq A} \mu_{xA} f_x(t)$$

$$f_C(t+dt) = f_C(t) - \mu_C f_C(t) + \sum_{x \neq C} \mu_{xC} f_x(t)$$

$$f_G(t+dt) = f_G(t) - \mu_G f_G(t) + \sum_{x \neq G} \mu_{xG} f_x(t)$$

$$f_U(t+dt) = f_U(t) - \mu_T f_U(t) + \sum_{x \neq U} \mu_{xU} f_x(t)$$

where $\mu_x = \sum_{y:y \neq x} \mu_{xy}$ is the rate of leaving state x. This process can be condensed into

$$\mathbf{F}(t+dt) = \mathbf{F}(t) + \mathbf{M}\mathbf{F}(t)dt$$

and therefore

$$\frac{d\mathbf{F}(t)}{dt} = \mathbf{M}\mathbf{F}(t) \tag{2.7}$$

where

$$\mathbf{M} = \begin{pmatrix} -\mu_A & \mu_{CA} & \mu_{GA} & \mu_{UA} \\ \mu_{AC} & -\mu_C & \mu_{GC} & \mu_{UC} \\ \mu_{AG} & \mu_{CG} & -\mu_G & \mu_{UG} \\ \mu_{AU} & \mu_{CU} & -\mu_{GU} & -\mu_U \end{pmatrix}.$$

The solution to the Equation 2.7 has the form (Yang, 2006)

$$\mathbf{F}(t) = e^{\mathbf{M}t}\mathbf{F}(0) \tag{2.8}$$

Thus, we have $\mathbf{P}(t) = e^{\mathbf{M}t}$ as the transition probabilities matrix with its entry $p_{xy}(t)$ being the transition probability at state y after evolutionary time t given the initial state x according to the model specified by the rate matrix $\mathbf{M}$.

2.2.3 Common models

The simplest form of a nucleotide substitution model(Jukes-Cantor model) assumes that the transition rates between any two nucleotide bases are the same (i.e., $\mu_{xy} = \mu$, $x \neq y$).

Under Jukes-Cantor model,

$$\mathbf{P}(t) = \begin{pmatrix} 1-3a_t & a_t & a_t & a_t \\ a_t & 1-3a_t & a_t & a_t \\ a_t & a_t & 1-3a_t & a_t \\ a_t & a_t & a_t & 1-3a_t \end{pmatrix}$$

where

$$a_t = \frac{1-e^{-4\mu t}}{4}. \tag{2.9}$$

Throughout we consider a Kimura model which allows different rates for transition and transversion. Let μ_1 and μ_2 be the rates of transition and transversion respectively. Under the Kimura model, a nucleotide base $x \in \omega_N$ mutates to a nucleotide base $y \in \omega_N$ within a time duration $\triangle t$ with probability

$$p_{\mu_1,\mu_2}(y|x, \triangle t) = 0.25 + 0.25e^{-4\mu_2\triangle t} + 0.5e^{-2(\mu_1+\mu_2)\triangle t}, \quad \text{for } x = y, \tag{2.10a}$$

$$p_{\mu_1,\mu_2}(y|x, \triangle t) \tag{2.10b}$$

$$= \begin{cases} 0.25 + 0.25e^{-4\mu_2\triangle t} - 0.5e^{-2(\mu_1+\mu_2)\triangle t}, & \text{for } x \neq y \text{ and it is a transition,} \\ 0.25 - 0.25e^{-4\mu_2\triangle t}, & \text{for } x \neq y \text{ and it is a transversion,} \end{cases} \tag{2.10c}$$

where μ_1 and μ_2 are the rates of transition and transversion respectively. Note that $\triangle t$

is arbitrary and does not have to be small for the equations above to hold. This process is quite general and not restricted to modelling only RNA virus mutations. More variants of nucleotide substitution models are explained in detail in Yang (2006).

2.2.4 Complexity of molecular evolution

Whilst Equation 2.10 is a common and mathematically tractable way to model molecular evolution, it is also an abstraction of a much more complex biological world. We discuss some aspects of molecular evolution for which Equation 2.10 does not account.

Molecular evolution is driven by both deterministic and stochastic forces, namely natural selection and genetic drift. If an *allele* [2] is more fit than other alleles in a particular environment, it is subject to *positive (natural) selection pressure*; similarly, it is subject to *negative selection pressure* if it is less fit. Deterministic evolution states that in an infinitely large population any other stochastic fluctuations in genomes, called *genetic drift*, do not affect the gene frequencies in the population. Hence, suppose the environmental factors and fitness of alleles are known, the evolutionary pattern is entirely predictable. In reality, genetic drift plays an important role in shaping the evolutionary pattern as the infinite population assumption rarely holds.

The *effective population size* [3] is the determining factor of the contribution of genetic drift to the evolution dynamics - i.e., the smaller effective population size, the larger the effect of genetic drift compared with the effect of natural selection. It is observed that when the population size varies across a few generations, the generation with least effective population size has a notable influence on the genetic diversity and hence the evolution pattern. As synonymous mutations do not alter functions of genes and are not subject to selection pressure, the effect of genetic drift may also be investigated by comparing synonymous and nonsynonymous mutation rates. *Neutral theory* (Kimura, 1984) suggests that genetic drift is the major force in shaping evolution as the effective population size in general is too small for positive selection to dominate. Equation 2.10 implicitly assumes a neutral model without accounting for any selection pressure.

Other underlying assumptions implied by Equation 2.10 may be questioned. For example, instead of having constant rates for all nucleotide sites, within a genome there can be several *conserved regions* (very low mutation rates) and *hypervariable regions* (very high mutation rates) (Yang, 1996). This, nevertheless, may be resolved

[2] An allele is one of the functional variants that can exist in particular positions on genomes.
[3] The effective size is the size of a randomly mating hypothetical population that has same allele frequencies as the observed population.

by introducing additional mutation parameters. Also, the time-reversible (Markov process) assumption may not be appropriate in some scenarios but non-reversible models are in general computationally difficult (Yang, 1994).

2.3 Bayesian inference

Having presented some common models for describing the epidemic and evolutionary processes, it is important, given observed data, to pursue statistical inference to make probability statements about the model parameters which are generally unknown.

From a Bayesian perspective, inference of the model parameter θ is entirely carried out on the *posterior distribution* of θ given the data y. Using Bayes' rule, the posterior distribution can be written as

$$\pi(\theta|y) \propto L(\theta; y)\pi(\theta) \tag{2.11}$$

where $L(\theta; y)$ is the likelihood function and $\pi(\theta)$ is the *prior distribution* of θ.

During epidemic outbreaks, data y typically represents only partial observations of the epidemic process which presents a challenging censoring issue. For example, the time of infection of an infected host is generally unavailable but also an essential component of the transmission dynamics − in many scenarios only the times of removal or the snapshots of infected sites are available (Ster et al, 2009; Lau et al, 2014b). It is now standard practice to conduct Bayesian inference of partially observed epidemics using the process of *data augmentation* supported by computational techniques such as Markov chain Monte Carlo methods (Neal and Roberts, 2004; Streftaris and Gibson, 2004b; Cook et al, 2007b; Gibson and Renshaw, 1998; O'Neill and Roberts, 1999). The details of these Bayesian techniques are given in this section following a discussion justifying the use of Bayesian over classical (frequentist) inference.

2.3.1 Why Bayesian?

The fundamental difference between Bayesian and classical inference is in their philosophical interpretations of the model parameter θ. In Bayesian inference θ is a random variable but in classical inference theory θ is an unknown fixed quantity.

A key result arising from this distinction in fundamental philosophy is that it makes sense only in Bayesian inference to talk about the *prior distribution* and *posterior*

distribution of θ which allows us to assign probabilities to θ taking values in a subset of the parameter space. For example [4], given observed data y, in Bayesian inference the quantity $P(\theta < 0.5|y)$ is a natural response to the question *whether* θ *is less than* 0.5. In contrast, classical inference requires to carry out a classical hypothesis testing at a certain significance level, say, 5%, to decide whether or not to reject the hypothesis in the question. However, the response from classical inference is much less meaningful as it assigns no probability (and uncertainty) to θ. It only says that if $\theta < 0.5$ the chance that the observed value of a certain test statistic $t(y)$ lies in a critical region is less than or equal to 5% - it is a probability statement about the data y instead of θ. The inability to make probability statements that apply directly to model parameters is a major limitation of classical inference. Yet from the practical perspective of modellers, it is often more important to be able to make probability statements about model parameters.

Bayesian inference theory is a coherent framework where the posterior distribution is derived purely with probability operations by combining the likelihood and the prior distribution. Classical inference on the other hand only utilises the likelihood and therefore lacks the machinery in Bayesian inference that can naturally combine the information on θ provided by the data (i.e., the likelihood) with the prior belief of θ from modellers (i.e., the prior distribution). With this machinery modellers may incorporate prior knowledge about some model parameters obtained from prior studies, which often facilitates the inference as the resolution of the data may not be sufficient to estimate these parameters.

Bayesian inference is also more favourable axiomatically. The *likelihood principle* requires that any inference should depend only on observed outcomes. However, many classical inference procedures (e.g., construction of confidence intervals, definition of consistency and measures of biasedness) rely on outcomes from the long-run behaviour of hypothetical repeated samplings. In contrast, Bayesian inference always conforms to this principle as the entire inference is based on the posterior distribution which depends on the data only through the likelihood function.

Interested readers are referred to O'Hagan et al (2004) for more in-depth discussion.

2.3.2 Noninformative prior

Although Bayesian framework provides the opportunity to make use of different sources of information in addition to the data, a main criticism against Bayesian

[4] An example in O'Hagan et al (2004).

inference is that the specification of the prior distribution can be subjective. A potential remedy is to use the so-called *noninformative priors*, although the meaning of *noninformative* can be vague, as we will see in the following discussion.

Flat prior

An obvious choice is to assign a flat (uniform) prior for θ. This approach, however, is not invariant to ways of parameterising θ. For example, let θ be the success probability of the Bernoulli distribution, we may specify a flat prior $U(0,1)$ for θ. Suppose we apply the transformation $h(\theta) = log(\frac{\theta}{1-\theta})$, the prior of $h(\theta)$ is apparently not a uniform distribution but becomes an informative prior (see Figure 2.1). As a result, two different parameterisations may lead to different posterior distributions given the same data.

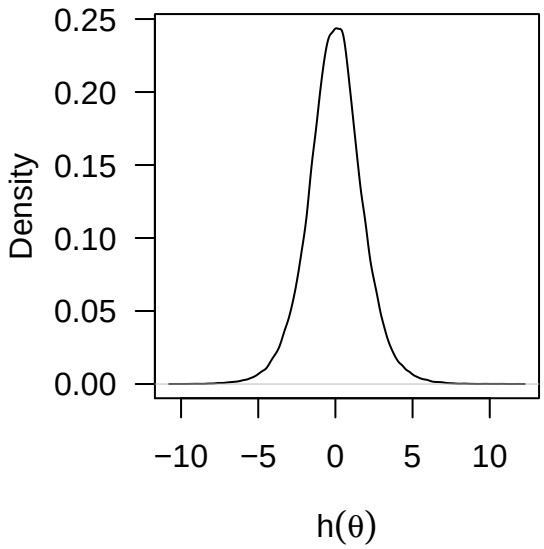


Figure 2.1: The density function of $h(\theta) = \frac{e^{h(\theta)}}{1+e^{h(\theta)}}$ where $\theta \sim U(0,1)$.

Jeffreys prior and reference prior

Jeffreys (1946) developed an approach to specify a prior that is invariant under reparameterisation. It is defined as

$$\pi(\theta) \propto \boldsymbol{I}(\theta)^{1/2} \tag{2.12}$$

with the *Fisher information* [5]

$$\boldsymbol{I}(\theta) = -\mathbb{E}_y\left[\frac{d^2 log(L(\theta; y)}{d\theta^2}\right] \tag{2.13}$$

where y is the data. Note that although a Jeffreys prior is considered to be noninformative, it is *not* necessarily flat on the parameter space, which can be counter-intuitive. For example, the Jeffreys prior for the success parameter in Binomial distribution is a $Beta(1/2, 1/2)$ distribution with much mass concentrated at two extremes of the interval $(0, 1)$. Although Jeffrey prior becomes flat when the Fisher information is constant according to Equation 2.12, the sense of Jeffreys prior being *noninformative* in general may become clearer after introducing *reference prior*. Roughly speaking, reference prior, proposed by Bernardo (1979) and followed by further development in Berger and Bernardo (1992), is a function that maximises the divergence between the posterior and prior given the data. By maximising the divergence, the data are then allowed to have the maximum effect on the posterior - this is the reason that the development of reference prior has been credited as a milestone of *objective Bayesian inference*. A common choice of divergence measure is the Kullback-Leibler (K-L) measure (Kullback, 1987) defined as

$$\int \pi(\theta|T = t) log \frac{\pi(\theta|T = t)}{\pi(\theta)} d\theta \tag{2.14}$$

where T is a sufficient statistic. Choosing a reference prior involves maximising the expectation over the distribution of T.

It can be shown that, for one-dimensional model parameter space, a reference prior is equivalent to a Jeffreys prior (Bernardo, 2005), which provides an intuition for the noninformative nature of Jeffreys prior. More intuition may be provided from the definition of the Fisher information − as Fisher information is the expected value of observed information, a prior derived from that may then be more data-driven and more objective.

Although flat priors suffer from the non-invariant property, they are commonly used in epidemic modelling when there is no strong prior information regarding model parameters (Cook et al, 2008; Kleczkowski and Gilligan, 2007). A main reason is that likelihood functions for disease dynamic models are generally not tractable and not differentiable which are required for deriving a Jeffrey or reference prior. The issue of intractable likelihood actually leads to the discussion in next subsection.

[5] Formally $log(L(\theta; y)$ should be written as $log(P(y|\theta))$ to reflect that y, instead of θ, is the random variable in this context.

2.3.3 Markov chain Monte Carlo (MCMC)

For a Bayesian, the whole spectrum of likelihood values on the parameter space is important for obtaining a complete picture of the posterior distribution. However, as we mentioned, likelihood functions for describing epidemic processes are generally not tractable. Fortunately, with the widespread availability of powerful computers, we can exploit a powerful but computationally intensive technique, called *Markov chain Monte Carlo (MCMC)*, which can be used to simulate samples from the targeted posterior distribution. This technique is utilised in Chapter 3 and Chapter 4.

Markov chain

A *Markov chain* is a discrete time stochastic process $\{\theta_0, \theta_1, \cdots\}$ where the next-time-step state θ_{t+1} at time $t+1$ is sampled with a *transition probability (kernel)* $P(\theta_{t+1} \in \Theta|\theta_t)$ conditional on the current state θ_t at time t, with Θ being the parameter space. The important implication is that, given the current state θ_t, the state θ_{t+1} is independent of the earlier history $\{\theta_0, \theta_1, \cdots \theta_{t-1}\}$.

Suppose the chain starts at state θ_0. Assuming *regularity conditions* (see later for more details) are satisfied, the distribution of θ_t will *eventually* converge to a *stationary distribution*. This property can be used to develop computational tools to implement Bayesian inference by constructing the Markov chain in such a way that the stationary distribution is the posterior distribution $\pi(\theta|y)$ of interest. Fortunately it turns out that is relatively straightforward if the posterior can be calculated up to a constant of proportionality (which is the case for most models of interest). Once the Markov chain has converged each iteration provides a sample from the posterior. Note that, although the chain is guaranteed to converge, there is no gold-standard rule to determine whether convergence has actually been achieved given a number of time steps. A common practice is to discard a sufficient number of initial samples (i.e., burn-in) and assess the convergence by visual inspection.

There are three regularity conditions to be satisfied to guarantee the convergence to the stationary distribution. The first condition requires the chain to be *irreducible*. That is, regardless of starting states, the chain must have positive probability of reaching any non-empty subset of Θ. Intuitively, this ensures that the parameter space can be thoroughly explored. The chain also needs to be *aperiodic*, which avoids the chain oscillating periodically between different states (Roberts, 1996). Lastly, the chain needs to be *positive recurrent* which requires that, in discrete state-spaces, the expected time to the first return to a state is finite. [6].

[6]See also other definitions for positive recurrence in (Roberts, 1996)

Metropolis-Hastings algorithm

Despite the seemingly strict regularity requirements, it turns out the construction of a Markov chain with a stationary distribution is surprisingly easy with the *Metropolis-Hastings* algorithm (Chib and Greenberg, 1995; Metropolis et al, 1953; Hastings, 1970). While key ideas of M-H algorithm are given in this section, interested readers are referred to an excellent textbook by Gilks (2005) which collects a good range of literature on this topic for more detailed discussion.

The M-H algorithm uses a transition kernel

$$P(\theta_{t+1}|\theta_t) = q(\theta_{t+1}|\theta_t) \times \alpha(\theta_t, \theta_{t+1}) \tag{2.15}$$

where $q(\theta_{t+1}|\theta_t)$ is the (conditional) *proposal distribution* from θ_t to θ_{t+1}, and $\alpha(\theta_t, \theta_{t+1})$ is the *acceptance probability* of this particular move, defined as follows:

$$\alpha(\theta_t, \theta_{t+1}) = \min\left\{1, \frac{\pi(\theta_{t+1}|y) \times q(\theta_t|\theta_{t+1})}{\pi(\theta_t|y) \times q(\theta_{t+1}|\theta_t)}\right\}. \tag{2.16}$$

It is straightforward to implement M-H algorithm: propose a next-step candidate θ' conditional on θ_t from $q(\theta'|\theta_t)$; let $\theta_{t+1} = \theta'$ with probability $\alpha(\theta_t, \theta')$, otherwise $\theta_{t+1} = \theta_t$. Note that this algorithm can be easily extended to multiple-parameter problems by updating the model parameters sequentially (i.e., conditioning on the current states of other model parameters when updating one particular parameter).

It can be easily shown that

$$\pi(\theta_t|y)q(\theta_{t+1}|\theta_t)\alpha(\theta_t, \theta_{t+1}) = \pi(\theta_{t+1}|y)q(\theta_t|\theta_{t+1})\alpha(\theta_{t+1}, \theta_t), \tag{2.17}$$

which then leads us to the *detailed balance equation*

$$\pi(\theta_t|y)P(\theta_{t+1}|\theta_t) = \pi(\theta_{t+1}|y)P(\theta_t|\theta_{t+1}). \tag{2.18}$$

Integrating both sides of Equation 2.18 with respect to θ_t, we obtain

$$\int \pi(\theta_t|y)P(\theta_{t+1}|\theta_t)d\theta_t = \pi(\theta_{t+1}|y) \tag{2.19}$$

which implies that if θ_t is from the stationary distribution $\pi(\cdot|y)$, θ_{t+1} is also from $\pi(\cdot|y)$. This actually fulfils the requirement of positive recurrence.

Proposal distribution In theory a proposal distribution can be of any form and the stationary distribution is invariant to that. However, in practice a good choice of

a proposal distribution is important, and it can be challenging for high-dimensional problems, for the rate of convergence and *mixing* of the chain (i.e., the frequency the chain moves around the mode of the stationary distribution). For example, in simplest cases, symmetric proposals such that $q(\theta_{t+1}|\theta_t) = q(\theta_t|\theta_{t+1})$ can be used [7]. In high-dimensional problems such as joint inference of epidemic and evolutionary process, the design of the proposal distribution requires much experimentation and creativity. In fact, a key innovation in Chapter 4 is to devise an efficient proposal for unobserved pathogen sequences circulating in the population.

Gibbs sampling θ_{t+1} may be sampled directly from its marginal distribution conditional on other model parameters, if it is known. This is called a *Gibbs sampling* (Casella and George, 1992) which is a special case of M-H algorithm where a move is always accepted. However, as for dynamic epidemic models, marginal conditional distributions of model parameters are in general not available analytically. O'Neill and Roberts (1999) demonstrate the use of Gibbs sampler in fitting the so-called Reed-Frost model which describe the spread of epidemics in a closed population in a discrete time frame (see also Bailey et al (1975)).

Other sampling methods Other sampling methods aiming at drawing samples from the posterior distribution are also available. Common non-Markov methods such as *rejection sampling* and *importance sampling*, requiring the identification of an envelope function or importance function which has a close relationship with the posterior distribution, can be difficult to implement in the context of epidemic models. *Particle filtering*[8] (Gordon et al, 1993) is another powerful tool becoming more popular as it offers the possibility of executing parallelised computer programs in which each particle can be generated independently, in contrast to MCMC where samples are dependent. Hybrid approaches that combine particle filtering and MCMC are recently developed (Andrieu et al, 2010) in which MCMC is conducted on parameter space and particle filtering is used to handle the nuisance parameters. Despite its great potential in improving the run-time of analysis and increasing popularity in epidemic modelling (Jégat et al, 2008; Ong et al, 2010; Skvortsov and Ristic, 2012), it still requires further developments particularly in high-dimensional problems. For example, it can be difficult to choose the number of particles and to have efficient moves of the particles in a high-dimensional space. Also, particle filtering tends to focus standard model parameters and is less concerned with augmented data such as infection events which are of key interest in epidemic modelling. Another approach

[7] This is commonly achieved by performing random-walk on the current state of the chain. For example, propose $\theta_{t+1} = \theta_t + N(0, 1)$.

[8] It can be considered as a recursive version of importance sampling.

called *Approximate Bayesian Computation (ABC)* (Tavaré et al, 1997; Diggle and Gratton, 1984) is gaining popularity in epidemiological and ecological studies (Neal, 2012; Beaumont, 2010; Tanaka et al, 2006), mainly due to its extreme simplicity. It is a likelihood-free approach, which samples parameters from the priors , simulates data from the model with these parameters and compares them with observed data − if the simulated data are too different from the observed data (difference is measured by a pre-defined discrepancy measure), the sampled model parameters are discarded and other retained samples will form the posterior distribution. However, the choice of the discrepancy measure can be subjective and the estimation can be inaccurate by simply comparing the observed data from a complicated dynamic system (see also section 2.5.3).

2.4 Partially observed epidemic process and data augmentation

Inference of an epidemic process is typically hindered by its partially-observed nature. Data augmentation techniques treat the unobserved data (e.g., the transition times between model compartments) as additional parameters and pursue the imputations of these quantities. In a simpler scenario where the number of transitions (e.g., number of exposed individuals) is known, the transition times may be sampled in the same manner as we discussed in previous sections. However, the sampling of the transition times is more complicated when the number of transitions is unknown. For instance, a typical unobserved component is the number of transitions to class E (i.e., the number of exposures). To handle this, a standard M-H algorithm described above is not sufficient as now the problem involves changes of the model dimension. A modified M-H algorithm, the *reversible jump Markov chain Monte Carlo (RJMCMC)* proposed by Green (1995) is required.

RJMCMC can be considered as a generalisation of M-H which can handle jumps between two models with different dimensions. Consider the simplest case where we want to add *one* transition to class E (i.e., adding one exposure) − for simplicity, we may denote this jump as $\theta \to (\theta_1, \theta_2)$ in which the proposed jump increases the model dimension by one. RJMCMC requires the "balance" of model dimension by generating one random variable u from a known density function $Q(u)$. (θ, u) is then mapped to (θ_1, θ_2) through a deterministic and invertible function h. The required proposal probability ratio $\frac{q\{(\theta,u)|(\theta_1,\theta_2)\}}{q\{(\theta_1,\theta_2)|(\theta,u)\}}$ is generally straightforward to compute. In addition, we need to compute a Jacobian term which accounts for the transformation carried out by the function h. In this simple example, we have the Jacobian term

as

$$\left| \begin{matrix} \frac{\partial \theta_1}{\partial \theta} & \frac{\partial \theta_1}{\partial u} \\ \frac{\partial \theta_2}{\partial \theta} & \frac{\partial \theta_2}{\partial u} \end{matrix} \right|. \tag{2.20}$$

The exact value of the Jacobian depends on the choice of h. The reversed jump (i.e., deleting an exposure) is similar by utilising the inverse function of h.

Gibson and Renshaw (1998) demonstrated the use of this algorithm for partially observed epidemics in the setting of *SEIR* models. For example, they allow a susceptible site moves to the set of exposed sites (i.e., a move from class S to class E) using the following algorithm:

(a) Randomly choose a site from the set of susceptibles ξ_U and move it to the set of exposed ξ_E. Uniformly draw an exposure time $E^{'}$ between $(0, t_{max})$ for this newly added exposure, where t_{max} is the upper bound of the exposure time.

(b) Denote n_u and n_E as the number of sites in current sets ξ_U and ξ_E respectively. Accept the proposed new sets and new exposure time with probability

$$\frac{L(\boldsymbol{\theta}; \mathbf{z}')}{L(\boldsymbol{\theta}; \mathbf{z})} \times \frac{n_u \times t_{max}}{1 + n_E} \tag{2.21}$$

where $\mathbf{z}'$ denotes the data with the changed sets of susceptibles and exposed and with the current exposure time E replaced by $E^{'}$. The second term of the acceptance probability is the proposal probability ratio and we have assumed that a reversed jump (i.e., deletion of an exposure) has an equal probability as this addition operation. The Jacobian equals 1 in this case.

More detailed description of reversible jump algorithms to epidemic models are given in later chapters.

We have so far presented basic ideas for some common models and statistical methods being used in epidemic modelling. These techniques are also required to develop and demonstrate the work in Chapter 3 and Chapter 4. In the next section, before we introduce the novel model assessment framework in the next chapter, we give an overview of the main Bayesian model selection techniques and highlight their limitations due to their complexity and sensitivity for epidemiological model assessments.

2.5 Bayesian model selection techniques

2.5.1 Bayes factor

A natural choice for model comparison is the Bayes factor invented by Jeffreys (1935, 1961) which formulates the model comparison problem into a hypothesis testing framework which indicates the *posterior* odds of the null model against the alternative model. The Bayes factor

$$B_{12} = \frac{P(\mathbf{D}|M_1)}{P(\mathbf{D}|M_2)} \tag{2.22}$$

is essentially the ratio of the marginal likelihoods with *observed data* $\mathbf{D}$ under two competing models M_1 and M_2, which provides a natural summary of evidence provided by the data in favor of M_1 against M_2. However, the implementation of the Bayes factor is largely hindered by the computation of the marginal likelihood

$$P(\mathbf{D}|M_i) = \int P(\mathbf{D}|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i \tag{2.23}$$

where θ_i is the model parameter under M_i. This integral is often intractable and can not be evaluated analytically (true for most epidemic models). Various methods for computing the Bayes factor have been proposed, including numerical evaluations, asymptotic approximation and approximation relying on simulating from the posterior (Kass and Raftery, 1995; Han and Carlin, 2001; Chib and Greenberg, 1995). Nevertheless, these methods are either prohibitive in terms of implementation or requiring approximations which might not be valid generally. Another issue is that the Bayes factor is known to be sensitive to the prior of the model parameter (i.e., $\pi(\theta_i|M_i)$ in Equation (2.23)), so one may draw a different conclusion regarding the model choice should different priors be adopted (Kass and Raftery, 1995). The computation of Bayes factor can be even more complicated when in the presence of unobserved data (Kass and Raftery, 1995), which is also often the case in epidemic modelling. On the other hand, a theoretically restrictive issue is the disagreement between the classical p-value and Bayes factor (i.e., they often draw conflicting conclusions in whether to accept the null) − an illustrative example of this conflict was given by Stone (1997).

2.5.2 Deviance Information Criterion (DIC)

Information theoretic approaches have also been considered. Examples include Laud and Ibrahim (1995), Gelfand and Ghosh (1998) and Spiegelhalter et al (2002). A popular approach, named *Deviance Information Criterion* (*DIC*), developed by Spiegel-

halter et al (2002) was claimed to be more applicable in comparing model adequacy of complex hierarchical models, for it reflected the complexity of a hierarchical model more genuinely.

The definition of DIC is based on a *deviance* $D(\theta) = -2logf(x|\theta) + 2logh(x)$, where $h(x)$ is a standardising term which is a function of data x. For model comparison, $h(x)$ is set to be 1 and we therefore have

$$D(\theta) = -2logf(x|\theta) \tag{2.24}$$

DIC (Spiegelhalter et al, 2002) is defined as follows:

$$DIC = \overline{D(\theta)} + p_D. \tag{2.25}$$

$\overline{D(\theta)} = \mathbb{E}_{\boldsymbol{\theta}}[log(f(x|\theta)|x]$ is called *posterior mean deviance* which is regarded as the Bayesian measure of fit. $p_D = \overline{D(\theta)} - D(\hat{\theta})$ is called the *effective dimension* of the model (where $\hat{\theta}$ is an estimate of the model parameter such as the posterior mean). The idea is that an over-fitted model with high dimension would be penalised by p_D. A smaller DIC corresponds to a better fit.

This approach relies on asymptotic approximations which are not necessarily statistically consistent with the underlying transmission mechanisms of epidemic models. More importantly, DIC is known to be problematic when applied to processes that are only partially observed as its definition becomes ambiguous (Celeux et al, 2006), as in the case of most real-world systems. As noted in Celeux et al (2006), missing data models can have many alternative representations of the DIC depending on the chosen representation for the missing data structure − particularly it depends on whether or not (or which) missing variables are identified as parameters in defining the model dimension and p_D. There are two main categories of DIC in the context of missing data models, namely the *complete DICs* and the *conditional DICs*, and among each category there are also variations. Denoting $\boldsymbol{y}$ and $\boldsymbol{z}$ and $\boldsymbol{\theta}$ as observed and unobserved data and the parameter vector respectively, complete DICs consider the complete likelihood $f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})$ and have a general form

$$DIC(\boldsymbol{y}, \boldsymbol{z}) = -4\mathbb{E}_{\boldsymbol{\theta}}[log(f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{z}] + 2log(f(\boldsymbol{y}, \boldsymbol{z}|\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}]). \tag{2.26}$$

The second category, conditional DICs, uses a different inferential focus and consider $\boldsymbol{z}$ as an additional parameter − instead of using the complete likelihood, it considers the conditional likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{z})$. Among each category the definition varies by relocating the position of the *log* and of the *expectations*. For example, in the category

of complete DICs, we can define

$$DIC_4 = -4\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[log(f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})|\boldsymbol{y}] + 2\mathbb{E}_{\boldsymbol{z}}[log(f(\boldsymbol{y},\boldsymbol{z}|\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}])|\boldsymbol{y}]. \tag{2.27}$$

A variation is to replace the second term in Equation 2.27 by taking out the expectation and replace $\boldsymbol{z}$ and $\boldsymbol{\theta}$ by some estimates (which are open to variations as well), which leads to

$$DIC_5 = -4\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{z}}[log(f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})|\boldsymbol{y}] + 2log(f(\boldsymbol{y},\hat{\boldsymbol{z}}|\hat{\boldsymbol{\theta}})). \tag{2.28}$$

The numbering of the versions of DICs refers to Celeux et al (2006). The performance of DIC at measuring model fit is explored in Chapter 3.

2.5.3 Posterior predictive checks

Denoting $\widetilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ as the predicted (observable) process and the actual observed process respectively, the *posterior predictive distribution* of $\widetilde{\boldsymbol{y}}$ is defined as

$$p(\widetilde{\boldsymbol{y}}|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\widetilde{\boldsymbol{y}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{2.29}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution of model parameters vector $\boldsymbol{\theta}$.

Rubin (1984) uses the posterior predictive distribution of a statistic to calculate the tail-area probability regarding to the observed value of the statistic. It is named as *posterior predictive p-value* and extended in Meng (1994). Conventional posterior predictive checking, which has roots in these papers, provides a common technique for epidemic models comparison (Cook et al, 2007b; Gibson et al, 2006). They compute summary statistics of the *observed* process simulated from competing models (with model parameters simulated from the posterior). This approach is relatively straightforward to implement and its results are very interpretable. However, it is in general difficult to identify sensitive summary statistics. Also, the main limitation of this approach is that these summary statistics only utilise the observed data, which may only reflect the averaging behaviour of models (this is demonstrated in Chapter 3).

Instead of relying on these *ad-hoc* summary statistics, in Chapter 3 we propose an innovative approach that first represents *exactly* the assumed epidemic process with residual processes that have a known distribution and independent of model assumptions. Subsequently, the posterior samples of these residual processes are imputed and they are tested against their theoretical distribution, from which we obtain a set of p-values as indications of model fits. We compare conventional posterior predictive

checks with our approach in Chapter 3.

Other approaches exist. For example, there have been efforts to address the conflict between classical and Bayesian hypothesis testing. Dempster (1974, 1997) proposed a likelihood ratio test using posterior distributions, and the method was extended by Aitkin (1997) and Aitkin et al (2005). However, they are dedicated to point null hypotheses, which render them too restrictive for general model selection problems in epidemiological studies where complicated non-nested models are often considered.

Chapter 3

A novel model assessment framework for spatio-temporal models in epidemiology and ecology

3.1 Introduction

As discussed in the introduction in Chapter 1, the predicted dynamics of dynamic spatio-temporal systems can be extremely sensitive to the choice of model, with consequent implications for the design of control strategies (Ster et al, 2009; Ferguson et al, 2001), but as yet there is a lack of effective model assessment tools described in the literature. At first glance Bayesian model selection techniques appear to be appealing. However, as discussed in 2.5, they suffer from key limitations due to their complexity and sensitivity for epidemiological model assessment.

Here we develop a novel approach for diagnosing mis-specifications of a general spatio-temporal transmission model by embedding classical ideas within a Bayesian analysis. Specifically, by proposing suitably designed non-centered parameterisation schemes, we construct latent residuals whose sampling properties are known given the model specification and which can be used to measure overall fit and to elicit evidence of the nature of mis-specifications of spatial and temporal processes included in the model. This model assessment approach can readily be implemented as an addendum to standard estimation algorithms for sampling from the posterior distributions such as Markov chain Monte Carlo. The proposed methodology is first tested using simulated data and subsequently applied to data describing the spread of *Heracleum*

mantegazzianum (giant hogweed) across Great Britain over a thirty-year period. The proposed methods are compared with alternative techniques including posterior predictive checking and the DIC. Results show that the proposed diagnostic tools are effective in assessing competing stochastic spatio-temporal transmission models and may offer improvements in power to detect model mis-specifications. Moreover, the latent-residual framework introduced here extends readily to a broad range of ecological and epidemiological models. We also extend the testing framework by developing a sequential procedure of the latent-residuals test. Results show that this sequential procedure, in contrast to the non-sequential approach, may exhibit higher sensitivity in scenarios when the observations in the early stage of the epidemic encapsulate more information on the transmission dynamics. Some of the results in this chapter are reported in Lau et al (2014b).

In this chapter we address the gap in available methodology by pursuing the following objectives:

- to innovate a statistically sound framework for assessing stochastic spatio-temporal models, which can be readily implemented as an addendum to a Bayesian analysis and which avoids the sensitivity and complexity of Bayesian model assessment;
- to illustrate how the approach can be targeted to assess particular aspects of a spatio-temporal stochastic model, here principally the choice of spatial kernel;
- to demonstrate the effectiveness of the approach using simulated data and to apply it to an ecological dataset describing the spread of an alien species throughout Great Britain.

The approach adopted (see Section 3.2) involves representing stochastic spatio-temporal models using appropriately designed *non-centered* parameterisation schemes (Papaspiliopoulos et al, 2007), from which *latent residual processes* can be defined. The assessment of fit of a model to a given data set, is then achieved in the Bayesian framework by imputing these residuals, and testing them for compliance with their (known) sampling model using classical tests. The approach has its roots in the framework proposed in Meng (1994), and extended in Streftaris and Gibson (2004a) and Gibson et al (2006). The key innovation in this work is to design the residual processes so that the resulting tests are sensitive to mis-specification of specific aspects of the model under consideration. In particular, we formulate tests for detecting mis-specification of the spatial transmission kernel as this aspect typically has major implications for control strategies, for example based on culling or removal of susceptibles in the neighbourhood of infected individuals.

3.2 Model and Methods

3.2.1 Spatio-temporal stochastic model

We consider a broad class of spatio-temporal stochastic models that we have discussed in 2.1. To facilitate reading of this chapter, here we recap some of the details. Throughout this chapter, we consider spatial SEIR epidemic model with susceptible (S), exposed (E), infectious (I) and removed (R) compartments. Suppose that we have a spatially distributed population indexed 1, 2, ..., denote $\xi_S(t)$, $\xi_E(t)$, $\xi_I(t)$ and $\xi_R(t)$ as the set of indices for individuals who are in class S, E, I and class R respectively at time t, and let $S(t)$, $E(t)$, $I(t)$ and $R(t)$ be the respective numbers of individuals in these classes at time t. Then individual $j \in \xi_S(t)$ becomes exposed during $[t, t+dt)$ with probability

$$p(j,t) = \{\alpha + \beta \sum_{i \in \xi_I(t)} K(d_{ij}, \kappa)\}dt + o(dt), \tag{3.1}$$

where α represents a primary infection rate and β is a contact parameter. We use a $Gamma(\mu, \sigma^2)$ parameterised by the mean, μ, and variance, σ^2, for the random time x spent in class E. For the random time x spent in class I we use a $Weibull(\gamma, \eta)$ parameterised by the shape and scale. All sojourn times are assumed independent of each other given the model parameters.

3.2.2 Latent residuals

Let $\mathbf{Z}$ denote the complete set of data (that is the time, nature and affected individual for all transitions between states) describing an epidemic generated randomly from the above model parametrised by $\boldsymbol{\theta}$. Then, as long as the sampling properties of the stochastic model described by Equation (3.1) are preserved, we can consider $\boldsymbol{Z}$ to be generated in non-unique ways. In particular, we consider $\boldsymbol{Z}$ as a deterministic function $h_{\boldsymbol{\theta}}(\cdot)$ of a random vector $\widetilde{\boldsymbol{r}}$ where the components of the latter are generated as a random sample from a Uniform(0, 1) distribution. That is

$$\boldsymbol{Z} = h_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{r}}). \tag{3.2}$$

This representation is essentially a *functional model* in the sense of Dawid and Stone (1982), and is an illustration of the concept of generalised residuals proposed in Cox and Snell (1968) (see Figure 3.1 for a schematic illustration). Note that the selection of a residual process $\widetilde{\boldsymbol{r}}$ and a function $h_{\boldsymbol{\theta}}(\cdot)$ which together specify the model given by Equation (3.1) can be effected in a multiplicity of ways. In the following sections we consider how this selection may be done in order to facilitate the design of statistical

tests based on $\widetilde{\boldsymbol{r}}$ that can be sensitive to mis-specification of particular aspects of the model.

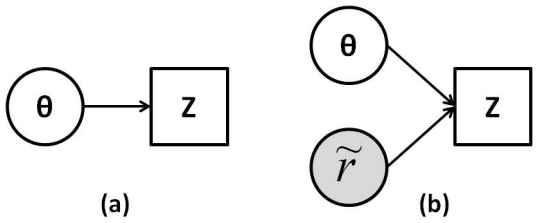


Figure 3.1: A graphical comparison between the centered and the non-centered parameterisation. (a) The centered parameterisation; (b) The non-centered parameterisation/ functional model representation.

The particular construction we exploit involves a process $\widetilde{\boldsymbol{r}}$ that can be partitioned into four independent random samples from Uniform(0, 1) and expressed as $\widetilde{\boldsymbol{r}} = (\widetilde{r}_1, \widetilde{r}_2, \widetilde{r}_3, \widetilde{r}_4)$ where each $\widetilde{r}_j$, $j = 1, 2, 3, 4$, is a vector which determines events relating to a different aspect of the process. The process $\widetilde{r}_1 = (r_{11}, r_{12}, ...)$ defines a set of population-level thresholds from which the time of each subsequent exposure can be determined by considering the integrated infectious challenge. For the k^{th} exposure (ordered temporally), r_{3k} and r_{4k} specify the quantile of the random sojourn times in class E and I respectively. The residuals $\widetilde{r}_2$ determine the particular infectious contacts that generate each exposure. We now describe concisely how the epidemic process can be constructed through the residuals $\widetilde{r}_1$, $\widetilde{r}_2$ and $\widetilde{r}_3$. A full description of the residuals can be found later (Section 3.3).

Exposure Time Residuals (ETR)

We refer to the residuals $\widetilde{r}_1$ as *Exposure Time Residuals (ETR)*. Starting from the $(k-1)^{th}$ exposure event, we define the accumulated infectious challenge in the population by time t as

$$A_{k-1}(t) = \int_{t_{k-1}}^{t} \sum_{j \in \xi_S(y)} \{\alpha + \beta \sum_{i \in \xi_I(y)} K(d_{ij}, \kappa)\} dy$$

where t_{k-1} is the time of the $(k-1)^{th}$ exposure event. The time of k^{th} exposure is then determined from

$$t_k = \inf \left\{ \, t \mid 1 - e^{-A_{k-1}(t)} > r_{1k} \right\}. \tag{3.3}$$

Infection-Link Residuals (ILR)

We refer to the residuals $\widetilde{r}_2$ as *Infection-Link Residuals (ILR)*. Given that the k^{th} exposure event occurs during $(t_k, t_k + dt)$, and given the other transitions that have occurred prior to t_k, the probability that the respective contact is between susceptible $j \in \xi_S(t_k)$ and infective $i \in \xi_I(t_k)$, is given by

$$p_{ij} \propto \beta K(d_{ij}, \kappa). \tag{3.4}$$

Note that the primary infection process can be accommodated by adding a notional and permanently infectious individual which presents a challenge α to every susceptible. To generate the particular infection-link from the residual, r_{2k}, we arrange the p_{ij} in the ascending order $p_{(1)}, \ldots, p_{(m)}$ where $m = S(t_k)(I(t_k)+1)$ is the total number of 'active' links. The link responsible for the k^{th} exposure is then determined from

$$s' = \inf \left\{ s \mid \sum_{a=1}^{s} p_{(a)} > r_{2k} \right\}, \tag{3.5}$$

so that individual j becomes exposed due to contact with individual i, and $p_{ij} = p_{(s')}$. The inclusion of the ordering operation in this process is motivated by our aim of designing tests that may be sensitive to mis-specification of the spatial kernel function in the model Equation (3.1). Suppose that the modelled kernel function deviates from reality in a systematic way − for example by declining too rapidly, or too slowly with distance. Then a heuristic argument (see later in Section 3.4.1) suggests that we may see a correspondingly systematic deviation from the U(0, 1) distribution when the residuals $\widetilde{r}_2$ are imputed from observations, with a concentration, or scarcity of residuals at the extremes, so that model mis-specification may be readily detected using standard tests of the fit of the imputed $\widetilde{r}_2$ to the uniform distribution.

Latent Time Residuals (LTR)

We refer to the residuals $\widetilde{r}_3$ as *Latent Time Residuals (LTR)*. For the k^{th} exposure, define the accumulated pressure of becoming infectious by time t as

$$Q(t) = \int_{t_k}^{t} \frac{f(y)}{1 - F(y)} dy, \tag{3.6}$$

where $f(y)$ and $F(y)$ are to the density and cumulative distribution function of the latent period respectively. The time of becoming infectious is then determined

from

$$t'_k = \inf\left\{ t \mid 1 - e^{-Q(t)} > r_{3k} \right\}. \tag{3.7}$$

Note that $1-e^{-Q(t)}$ equivalently gives the quantile of the cumulative distribution function of the latent period and hence the time becoming infectious can be computed by using usual inversion of the cumulative distribution function (see also Section 3.3.3). We remark that the time of recovery can be determined similarly by utilizing r_{4k} and an appropriate sojourn time distribution in class I.

3.3 Reconstructing the epidemic using the residual process

In this section, we provide the theoretical justification for the use of latent residuals in reconstructing the epidemic process. As we shall see, three independent random draws of $U(0,1)$ variates are involved in constructing three different aspects of the epidemic process.

3.3.1 Transition probabilities

We first formulate the targeted transition probabilities that our reconstruction scheme with the residual process should aim at. A transition event from S to E (i.e., an infection event) in time interval $(t, t+dt)$ is characterised by the following probability,

$$\begin{aligned} &P(S(t+dt) = s-1 | S(t) = s) \\ &= \{ \sum_{j \in \xi_S(t)} (\alpha + \beta \sum_{i \in \xi_I(t)} K(d_{ij}, \kappa)) \} dt + o(dt). \end{aligned} \tag{3.8}$$

The transition of an exposed individual j from class E to class I in the time interval $(t, t+dt)$ is governed by the following probability,

$$P(j \in \xi_I(t+dt) | j \in \xi_E(t)) = h_T(t)dt + o(dt), \tag{3.9}$$

where $h_T(\cdot)$ is the hazard rate function corresponding to the waiting time T from class E to class I. The transition form class I to class R should follow an analogous form given a waiting time distribution between these two classes.

3.3.2 Sellke thresholds, construction of the exposure times and the infection links

As discussed in 2.1.2, Sellke (1983) demonstrated the equivalence of a threshold model and a standard time-homogeneous Markov process if the thresholds are assumed to be independent and identically distributed exponential random variables with mean equal to 1. The key assumption in this threshold model is that an individual possesses a *random resistance to infection* which is termed as the *Sellke threshold*. That is, the individual would only be infected when the *infective pressure* (see Equation 2.5) from the infectious reaches the Sellke threshold assigned. Sellke's construction requires a threshold to be assigned to each individual in the population, and therefore censoring of the threshold of a non-exposed individual is necessary (Gibson et al, 2006).

To avoid the censoring issue in using Sellke thresholds, to minimise the number of residuals that must be imputed in practice, and to allow for the construction of a test specifically for testing the fit of the spatial kernel, we consider a construction by defining a population-level threshold for an infection event (in contrast to the individual-level threshold in the Sellke construction) and explicitly construct the infection links within the *competing-risk framework* (i.e., a failure event could only occur in one and only one of the possible ways to fail). We first define the *accumulated infectious challenge* $G(t)$ in the population by time t as

$$G(t) = \int_0^t \sum_{j \in \xi_S(y)} \{\alpha + \beta \sum_{i \in \xi_I(y)} K(d_{ij}, \kappa)\} dy. \tag{3.10}$$

Also define the *threshold* r_{1k} (i.e., ETR) for the k^{th} infection event as

$$r_{1k} \sim Exp(1).$$

Note that r_{1k} can be transformed to $U(0,1)$ by the cumulative distribution function of the exponential distribution. It is also noted that the threshold we defined is on the population-level (i.e., aggregating all infective challenges in the population) instead of assigning an individual threshold for each susceptible in the population. Lastly we let $h_G(\cdot)$ be the hazard function of the random threshold r_{1k}. We then prove the mechanism defined below is equivalent to the mechanism specified by equation (3.8).

Proposition 3.3.1 *A threshold model, which states that the k^{th} infection event would occur at time t, where $t \geq t_{k-1}$ and $t_0 = 0$, only if $A_{k-1}(t) \geq r_{1k}$ (i.e., $t = \inf\{ t' \mid A_{k-1}(t') > r_{1k}\}$), is equivalent to the mechanism specified by equation* (3.8).

Proof Denoting $\mathcal{F}_t$ as the history of the change of infectiousness before time t, the probability of having the k^{th} infection event at time interval $(t, t+dt)$, where $t \geq t_{k-1}$, from the threshold model defined in proposition 3.3.1 is given by the following,

$$\begin{aligned}
&P(A_{k-1}(t) \leq r_{1k} \leq A_{k-1}(t) + dA_{k-1}(t) | r_{1k} > A_{k-1}(t), t_{k-1}, \mathcal{F}_t) \\
&= h_G(A_{k-1}(t)) dA_{k-1}(t) + o(dA_{k-1}(t)) \\
&= dA_{k-1}(t) + o(dt) \quad (\because h_G(\cdot) = 1) \\
&= dG(t) + o(dt) \\
&= \sum_{j \in \xi_S(t)} \{\alpha + \beta \sum_{i \in \xi_I(t)} K(d_{ij}, \kappa)\} dt + o(dt).
\end{aligned}$$

The second last equality holds as

$$A_{k-1}(t) = G(t) - G(t_{k-1}). \quad \blacksquare$$

Further, conditional on the occurrence of the k^{th} infection event at time t_k, we can construct the corresponding infection link within the competing-risk framework. We first denote p_{ij} as the probability of individual i infecting individual j where $i \in \xi_I(t_k)$ and $j \in \xi_S(t_k)$. By noting that $e_{ij} = z_{ij}dt$, where $z_{ij} = \alpha$ when considering the primary infection and $z_{ij} = \beta K(d_{ij})$ for $i \in \xi_I(t_k)$ and $j \in \xi_S(t_k)$, it can be readily seen that the total probability of having this infection event is in fact the sum of all e_{ij} and this sum is the same as the transmission probability in equation (3.8). Hence the equivalence is not altered in constructing the infection link corresponding to this infection event. The actual infection link is then determined by a random draw from $U(0,1)$ and the values of z_{ij}: we first sort all the $e'_{ij} = z_{ij} / \sum_{i,j} z_{ij}$ in ascending order and denote them as $e'_{(1)}, \ldots, e'_{(m)}$ where m is the total number of the possible links; we then draw a random number, r_{2k}, from $U(0,1)$ (i.e., the ILR), if $r_{2k} \in (\sum_j^{n-1} e'_{(j)}, \sum_j^n e'_{(j)})$ n^{th} link is realised as the actual infection link.

3.3.3 Construction of the sojourn time

We now propose a threshold model for the construction of the sojourn time in class E (i.e., the latent period) and prove its equivalence with the mechanism defined in

equation (3.9). Define the *threshold* r_{3k} (i.e., LTR) such that

$$r_{3k} \sim Exp(1).$$

Similarly, r_{3k} can be transformed to $U(0,1)$ by the cumulative distribution function of the exponential distribution. Finally define $h_Q(\cdot)$ to be the hazard function of the (random) threshold r_{3k}.

Proposition 3.3.2 *A threshold model, which states that the transition of an exposed individual k from class E to class I would occur at time t only if $Q(t) \geq r_{3k}$ (i.e. $t = \inf\{\, t' \mid Q(t') \geq r_{3k}\}$), is equivalent to the mechanism specified by equation* (3.9).

Proof The transition probability from class E to class I during time interval $(t, t+dt)$ from the threshold model defined in proposition 3.3.2 is given by the following,

$$\begin{aligned}
&P(Q(t) < r_{3k} \leq Q(t) + dQ(t) | k \in \xi_E(t), r_{3k} > Q(t)) \\
&= h_Q(Q(t))dQ(t) + o(dQ(t)) \\
&= dQ(t) + o(dt) \quad (\because h_Q(\cdot) = 1) \\
&= h_T(t)dt + o(dt) \quad \blacksquare
\end{aligned}$$

Denote $F_Q(\cdot)$ and $F_T(\cdot)$ as the cumulative distribution functions for the threshold and for the sojourn time in class E respectively. It should be noted that $F_T(t) = F_Q(Q(t))$ where $F_Q(Q(t)) \sim U(0,1)$. Therefore, the sojourn time can be obtained by computing $F_T^{-1}(r')$ where $r' \sim U(0,1)$.

It can be readily seen that this approach can be extended to transition between two classes in which the forms of sojourn time distributions are explicit (e.g., sojourn time in class I).

3.3.4 Detailed algorithm for simulating epidemics utilizing the residual process

We give the detailed algorithm for simulating the exposure times and infection links by utilizing the residual process. A realisation of an infection/exposure event can be summarised into two major steps − first the time of next infection is determined; then according to the probabilities e'_{ij} at a particular time of infection event, one of

infection links will be realised. The details of the realisation process of the epidemic can be performed in the following way:

1. Given the $(k-1)^{th}$ infection event, realise k^{th} infection event as follows.
2. Draw a threshold g'_k, from $Exp(1)$.
3. Given $\mathcal{F}_t$ for any $t > t_{k-1}$, compute t_k, the time of k^{th} infection event, by solving following equation for t

$$g'_k = \int_{t_{k-1}}^{t} \sum_{j \in \xi_S(t)} \{\alpha + \beta \sum_{i \in \xi_I(t)} K(d_{ij})\} dt.$$

4. Consider the system at time t_k, compute the normalised probabilities e'_{ij} according to the expression $e'_{ij} = z_{ij} / \sum_i \sum_j z_{ij}$.
5. Sort all e'_{ij} in an ascending order; sub-divide the interval [0,1] into sub-intervals whose widths are equal to the sorted elements and the ranks are preserved; also, record the rank for each link in the sorting.
6. Draw a random number r_k from $U(0,1)$.
7. If r_k lies in m^{th} interval, m^{th} infection link is realised.
8. Simulate the next infection event by repeating the above steps; continue the simulation until $t_k > t_{max}$, where t_{max} is the observation period of the epidemic.

The realisation of an epidemic in this way not only gives the equivalent sampling properties as given by the construction from Sellke thresholds, it also explicitly constructs the infection network. It requires only a sequential draw of thresholds and this allows us to avoid the censoring of the Sellke thresholds of non-infected individuals. The simulation of transition time from class E to class I can be performed in a similar manner by using the corresponding form of accumulated pressure in Equation 3.6.

3.3.5 Bayesian inference and model assessment

It is now standard practice to conduct Bayesian analyses of partially observed epidemics using the process of *data augmentation* supported by computational techniques such as Markov chain Monte Carlo methods described in detail in Chapter 2 (Ferguson et al, 2001; Catterall et al, 2012; Cook et al, 2007b; Gibson et al, 2006) . To recap, given partial data y, these approaches involve simulating from the joint posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{z}|y)$ where $\boldsymbol{z}$ represents the complete epidemic data as above. This approach, as applied to fit models in this chapter, is described more fully in

Sections 3.3.7 and 3.3.8. As the complete epidemic $\boldsymbol{z}$ is reconstructed, or 'imputed', it naturally lends itself to the residual-based testing methods now described.

Given a random draw $(\boldsymbol{\theta}', \boldsymbol{z}')$ from $\pi(\boldsymbol{\theta}, \boldsymbol{z}|y)$ it is generally straightforward to invert Equation (3.2) to impute the corresponding residual $\widetilde{\boldsymbol{r}}'$ by sampling it uniformly from interval obtained from the set $h^{-1}_{\boldsymbol{\theta}'}(\boldsymbol{z}')$, the set of residuals mapped to $\boldsymbol{z}'$ by $h_{\boldsymbol{\theta}'}$. Therefore a sample from the posterior distribution $\pi(\widetilde{\boldsymbol{r}}|y)$ can easily be generated with a minor insert to an existing algorithm. On applying a classical test (see below) for consistency with the uniform distribution to the $\widetilde{\boldsymbol{r}}'$ a posterior distribution of *p-values*, $\pi(P(\widetilde{\boldsymbol{r}})|y)$ is generated, from which evidence against the modelling assumptions can be discerned. In Bayesian parlance we note that the pair $(\boldsymbol{\theta}', \widetilde{\boldsymbol{r}}')$ represents a *non-centered parameterisation.*

Hypothesis testing Specifically we use an *Anderson-Darling* hypothesis test (Lewis, 1961) which has the test statistic

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[(\ln Y_{(i)} + \ln(1 - Y_{(n-i+1)})], \tag{3.11}$$

where n is the sample size and $Y_{(i)}$ is the i^{th} largest sample. The form of A is complicated and its percentiles have to be determined by numerical integration, saddlepoint or other approximation methods (Marsaglia and Marsaglia, 2004). Marsaglia and Marsaglia (2004) used extensive simulations to compute the percentiles and appropriate p-value. These computational tools are available in a package provided by the statistical software R (Bellosta, 2011), which we have used to compute the p-value. In view of our aim of detecting an anticipated mis-match in the tails of the distribution of imputed residuals, we do not adopt the commonly used Kolmogorov-Smirnov test which is known to be non-sensitive to the tail of the distribution.

Imputation of Infection-Link Residuals (ILR)

The imputation of $\widetilde{r}_1$ and $\widetilde{r}_3$ given $(\boldsymbol{\theta}', \boldsymbol{z}')$ is straightforward by inverting the procedure specified by Equation (3.3) and Equation (3.7) respectively (see also Section 3.3.4). Imputation of r_{2k} is achieved by inverting Equation (3.5) but, since the infection link is discrete and the space of residuals continuous, the imputation process warrants description here. The particular infection link for the k^{th} exposure event is randomly chosen from the links between the corresponding exposed individual k and $i \in \xi_I(t)$ according to probabilities p_{ik} defined in Equation (3.4). The ranking of this particular infection link, s', is then determined among all links between $j \in \xi_S(t)$ and infective

$i \in \xi_I(t)$. Finally, the residual r'_{2k} is imputed as a random draw

$$r'_{2k} \sim U\ (\sum_{i=1}^{i<s'} p_{(i)}, \sum_{i=1}^{i<s'} p_{(i)} + p_{(s')}). \tag{3.12}$$

Bayesian residuals have been utilised in other contexts (Albert and Chib, 1993). In Gibson et al (2006) it was shown that Bayesian latent residuals based on Sellke thresholds (Sellke, 1983) could be used to assess the fit of spatio-temporal models. However, as we have discussed previously, the specific approach is problematic when the epidemic is small as thresholds must be imputed even for uninfected individuals. By contrast the construction proposed here requires residuals to be imputed only for each infection event and avoids this shortcoming. Moreover, since the components of the residual process $\widetilde{\boldsymbol{r}}$ each relate to a different aspect of the stochastic model, then it is plausible that testing for mis-specification of a given aspect may be best achieved by considering only the relevant component of $\widetilde{\boldsymbol{r}}$. In particular, mis-specification of a spatial kernel or the latent period distribution may be assessed by examining the posterior samples of $\widetilde{r}_2$ (ILR), and $\widetilde{r}_3$ (LTR) respectively. We stress again the importance for the detection of mis-specified spatial kernels of the ordering operation in the construction of the ILR, which is included with the expectation that it leads to systematic, detectable, and interpretable deviations from U(0, 1) in the imputed residuals. This issue is discussed further in following sections. As described in Section 3.4 and Section 3.6, this hoped-for sensitivity is indeed achieved.

3.3.6 Interpretation of latent-residual tests

Posterior distributions of p-values arising from a classical test applied to a latent process have been exploited in (Gibson et al, 2006; Streftaris and Gibson, 2004a, 2012). For completeness we include some comments on the statistical interpretation of such distributions in the Bayesian context. To the Bayesian observer of data y, $\pi(P(\widetilde{\boldsymbol{r}})|y)$ represents their posterior belief regarding the p-value that a classical observer of $\widetilde{\boldsymbol{r}}$ would compute. Should this distribution be concentrated on small values, the Bayesian would infer that the classical observer may reject the hypothesis that the $\widetilde{\boldsymbol{r}}$ were generated as a random sample from a $U(0,1)$ distribution. The latter hypothesis is a key assumption for the functional-model representation given in Equation (3.2) so that the classical observer would likewise question the validity of this model. Therefore, the Bayesian observer can extract from $\pi(P(\widetilde{\boldsymbol{r}})|y)$ summaries such as $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|y)$ (as used here) and interpret them as measures of evidence against their model assumptions.

3.3.7 Likelihood

Consider a population of size N. Assume that individuals in the population are all susceptible at time 0 which is the time of introduction of the force of primary infection, and assume that the epidemic is to be observed up to time t_{max}. Let $\mathbf{E} = (E_1, E_2, \ldots, E_{N_E})$ be a vector of the exposure times of N_E individuals, $\mathbf{I} = (I_1, I_2, \ldots, I_{N_I})$ be a vector of the times of becoming infectious of N_I individuals; $\mathbf{R} = (R_1, R_2, \ldots, R_{N_R})$ be a vector of the times of recovery of N_R individuals. Also, we let χ_U be the set of indices of the individuals remaining in class S at the end of observation period t_{max}; and, to recap, we let χ_E, χ_I and χ_R be the set of indices of the individuals who have gone through class E, class I and class R by t_{max} respectively. Also, we let $\chi_{E\setminus I}$ be the set corresponding to the exposed individuals who have not been infectious up to time t_{max} and $\chi_{I\setminus R}$ be the set corresponding to the infectious individuals who have not recovered up to time t_{max}. The cumulative density functions for waiting times in class E and class I are denoted by $F_E(\cdot)$ and $F_I(\cdot)$ respectively. Finally, let $\boldsymbol{\theta} = (\alpha, \beta, \mu, \sigma^2, \gamma, \eta, \kappa)$ be vector of parameters in the model. As a result, we can express the likelihood function given the times of events as

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{E}, \mathbf{I}, \mathbf{R}) = &\{ \prod_{j \in \chi_E^{-1}} \{\alpha + \sum_{i \in \chi_{I\setminus R}(E_j{}^-)})\beta K(d_{ij}; \kappa)\} e^{-q'_j} \} \times \prod_{j \in \chi_U} e^{-q_{T_j}} \\ &\times \prod_{j \in \chi_I} f_E(I_j - E_j; \mu, \sigma^2) \times \prod_{j \in \chi_R} f_I(R_j - I_j; \gamma, \eta) \\ &\times \prod_{j \in \chi_{E\setminus I}} (1 - F_E(t_{max} - E_j; \mu, \sigma^2)) \times \prod_{j \in \chi_{I\setminus R}} (1 - F_I(t_{max} - I_j; \gamma, \eta)) \end{aligned} \tag{3.13}$$

where $I(t)$ is number of individuals in class I at time t, and ξ_E^{-1} is the set of individuals in class E excluding the index case. Also,

$$q'_j = \int_{t=0}^{E_j} \{\alpha + \sum_{i \in I_{t^-}} \beta K(d_{ij}; \kappa)\} dt, \tag{3.14}$$

and

$$q_{T_j} = \int_{t=0}^{t_{max}} \{\alpha + \sum_{i \in I_{t^-}} \beta K(d_{ij}; \kappa)\} dt. \tag{3.15}$$

Note that $e^{-q'_j}$ and $e^{-q_{T_j}}$ correspond to the likelihood of an exposure time and an

non-exposed individual respectively. The second and third lines in Equation 3.13 represent the contribution to the likelihood of the sojourn times in class E and I respectively.

3.3.8 Estimation

MCMC methods are employed to estimate the joint posterior. In particular we use the (single-step) Metropolis-Hastings algorithm (Chib and Greenberg, 1995) and update the model parameters sequentially. We assume uniform priors for the parameters. To allow cryptic exposures in the simulation study, following Gibson & Renshaw (Gibson and Renshaw, 1998), we adapt the reversible-jump algorithm (Green, 1995) to the compartmental model setting. We have discussed the general details of MCMC and in particular RJMCMC in Chapter 2. Details of specific application to our problem in this chapter are given below.

Single-step MH algorithm for model parameters

We first arbitrarily choose $\boldsymbol{\theta}_1$ and $\boldsymbol{E}_1$ such that $\pi(\boldsymbol{\theta}_1|\boldsymbol{E}_1, \boldsymbol{I}, \boldsymbol{R}) > 0$. Then we repeat the following steps for $i = 1, \ldots, n$ where n is the number of iterations.

I **Update α, β, κ, μ, σ^2, γ and η sequentially**

(a) Propose a new parameter value, α', by performing a random-walk on the corresponding current value of the parameter, α_i. Specifically, we have

$$\alpha' \sim N(\alpha_i, \sigma_1^2) \tag{3.16}$$

where we set $\sigma_1^2 = 1$. If $\alpha' < 0$, it is rejected and the current value is retained.

(b) Accept the proposed α' with probability

$$\frac{L(\boldsymbol{\theta}'; \boldsymbol{E}_i, \boldsymbol{I}, \boldsymbol{R})}{L(\boldsymbol{\theta}_i; \boldsymbol{E}_i, \boldsymbol{I}, \boldsymbol{R})} \tag{3.17}$$

where $\boldsymbol{\theta}'$ denotes the vector of parameters with α_i replaced by α'. Note that since we have used uniform priors and a symmetric proposal distribution, the acceptance probability reduces to the ratio of likelihoods.

(c) If α' is accepted, set $\alpha_{i+1} = \alpha'$, otherwise $\alpha_{i+1} = \alpha_i$. $\boldsymbol{\theta}_{i+1}$ is also updated accordingly.

(d) Apply the same algorithm to the remaining parameters sequentially.

II **Update the exposure times** E_j

(a) Randomly choose an exposure, j, and draw a new exposure time $E_j^{'}$ uniformly between $(0, t)$, where $t = I_j$ if j has become infectious, otherwise $t = t_{max}$.

(b) Accept the proposed new exposure time with probability

$$\frac{L(\boldsymbol{\theta}_{i+1}; \boldsymbol{E}', \boldsymbol{I}, \boldsymbol{R})}{L(\boldsymbol{\theta}_{i+1}; \boldsymbol{E}_i, \boldsymbol{I}, \boldsymbol{R})} \tag{3.18}$$

where $\boldsymbol{E}'$ denotes the data with the current exposure time E_j replaced by $E_j^{'}$.

(c) If accepted, $\boldsymbol{E}_{i+1} = \boldsymbol{E}'$ otherwise $\boldsymbol{E}_{i+1} = \boldsymbol{E}_i$.

Reversible jump algorithm for cryptic exposures

Individuals/sites that have been exposed but have not yet become infectious are referred to as cryptic exposures. In the simulation study (see Section 3.4) in which an SEIR model is fitted, we allow (unobserved) cryptic exposures and 'swap' of sites between the set $\xi_{E\setminus I}$ and ξ_U. As discussed in Chapter 2, these operations involve changes of model dimension, which requires the use of the reversible jump algorithm. Adapting from the methodology in Gibson and Renshaw (1998), we apply two operations an *addition* and a *deletion* on the set of $\xi_{E\setminus I}$. At each iteration during the MCMC run (following the updates of $\boldsymbol{\theta}$ and $\boldsymbol{E}$), each operation (deletion or addition) is equally likely to be applied.

I Addition of a cryptic exposure

(a) Randomly choose a site from ξ_U and move it to the set of $\xi_{E\setminus I}$ and ξ_E. Uniformly draw an exposure time $E_j^{'}$ between $(0, t_{max})$ for this newly added cryptic exposure.

(b) Denote n_u and $n_{E\setminus I}$ as the number of sites in current sets ξ_U and $\xi_{E\setminus I}$ respectively. Accept the proposed new sets and new exposure time with probability

$$\frac{L(\boldsymbol{\theta}; \mathbf{z}')}{L(\boldsymbol{\theta}; \mathbf{z})} \times \frac{n_u \times t_{max}}{1 + n_{E\setminus I}} \tag{3.19}$$

where $\mathbf{z}'$ denotes the data with the changed sets ($\xi_{E\setminus I}$ and ξ_E and ξ_U) and with the current exposure time E_j replaced by $E_j^{'}$.

II Deletion of a cryptic exposure

(a) Randomly choose a site from $\xi_{E\setminus I}$ (also from ξ_E) and move it to the set of ξ_U; delete the corresponding E_j accordingly.

(b) Accept the proposed new sets with probability

$$\frac{L(\boldsymbol{\theta}; \mathbf{z}')}{L(\boldsymbol{\theta}; \mathbf{z})} \times \frac{n_{E\setminus I}}{(1 + n_u) \times t_{max}} \tag{3.20}$$

where $\mathbf{z}'$ denotes the data with the changed sets ($\xi_{E\setminus I}$ and ξ_E and ξ_U) and with the current exposure time E_j being deleted.

3.4 Simulated example

To test the methodology we apply it to analyse spatio-temporal epidemics simulated in a population of size $N = 1000$, whose locations are sampled independently from a uniform distribution over a square region, between times $t = 0$ and $t = t_{max} = 50$. We assume that the entire population is susceptible at time 0, that the epidemic evolves according to Equation (3.1) with $\alpha = 0.001$, $\beta = 3$, $K(d_{ij}, \kappa_1) = \exp(-0.03 d_{ij})$, and that the sojourn times in classes E and I follow $Gamma(5, 2.5)$ and $Weibull(2, 2)$ distributions respectively. The observations y constitute only the precise times and locations of transitions from E to I and from I to R that occur during the observation period. Figure 3.2 illustrates the spatio-temporal progression of a typical realisation of y.

To assess our model testing framework we fit to the simulated data y a model with the correct structure (Case A), and three further models in which the spatial kernels (Case B & Case C) and the latent period distribution (Case D) have been mis-specified respectively. Specifically we consider:

- Case A: $K(d_{ij}, \kappa_1) = exp(-\kappa_1 d_{ij})$ and the latent period is distributed as Gamma(μ, σ^2);
- Case B: $K(d_{ij}, \kappa_2) = 1/(1 + d_{ij}/\kappa_2)$ and the latent period is distributed as Gamma(μ, σ^2);
- Case C: $K(d_{ij}, \kappa_3) = d_{ij}^{-\kappa_3}$ and the latent period is distributed as Gamma(μ, σ^2);
- Case D: $K(d_{ij}, \kappa_1) = exp(-\kappa_1 d_{ij})$ and the latent period is distributed as Exp(μ).

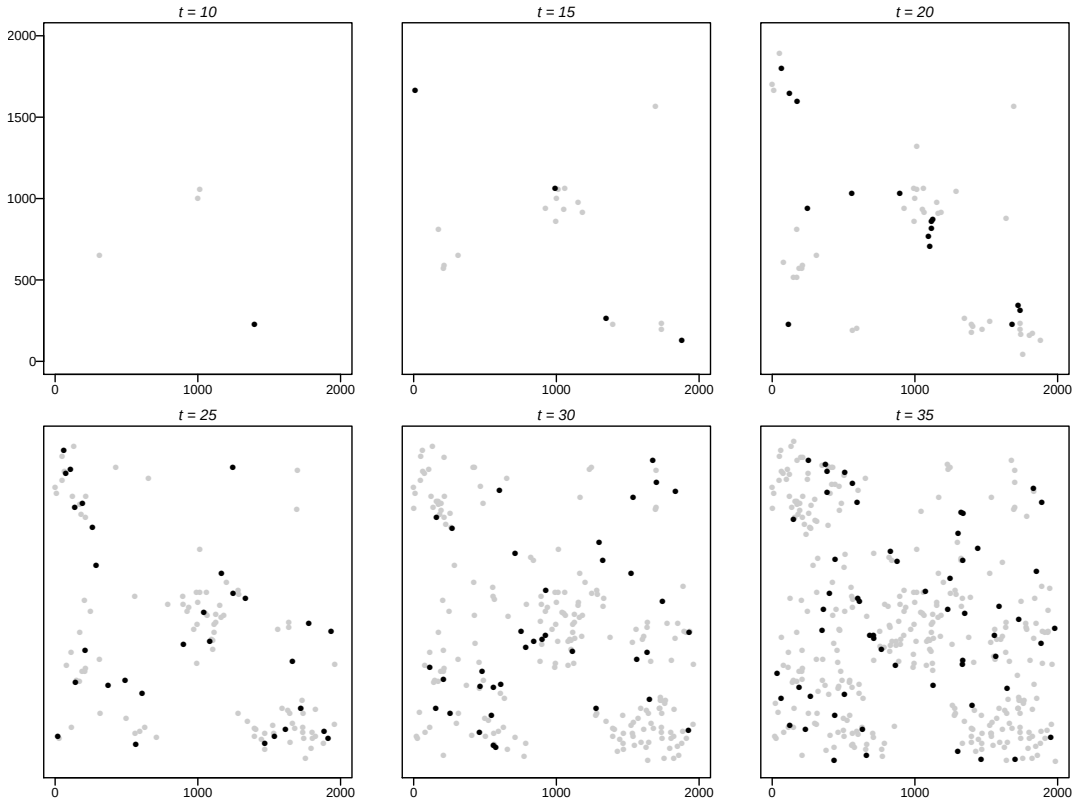


Figure 3.2: An illustration of (a subset of) the observed data y in the simulation (replicate 1) in the 2000×2000 square area in the form of a sequence 'snapshots' of the system state at particular times. Black and grey dots represent the individuals in class I and R respectively at times $t = 10, 15, 20, 25, 30, 35$. It is assumed that the locations of all other individuals are known but these are not shown in the interests of clarity.

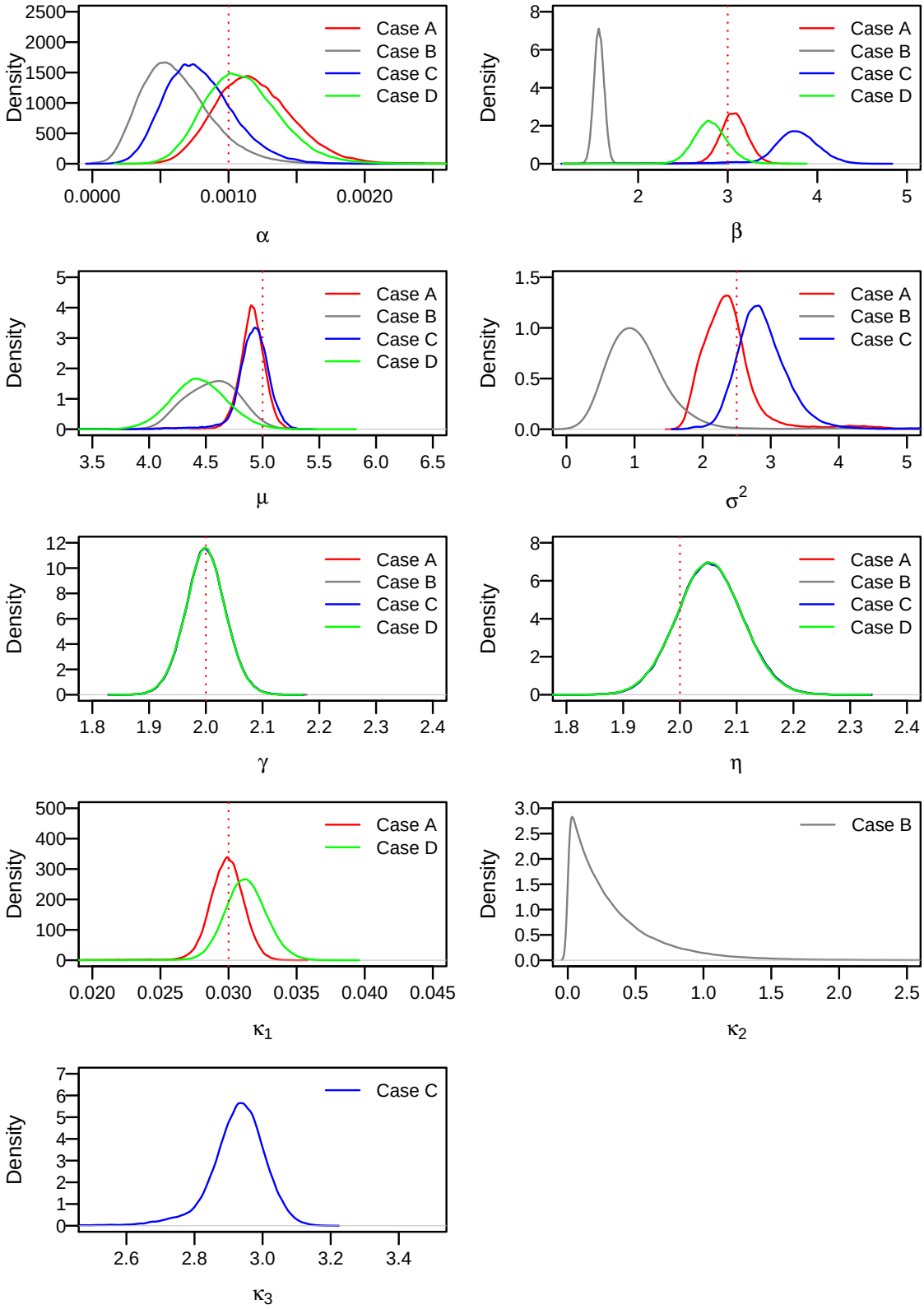


Figure 3.3: Posterior distributions of model parameters for models fitted to the simulated data (Replicate 1). Dotted lines represent the actual values of the parameters used for simulating the epidemic. Note that the posterior distributions of the model parameters γ and η corresponds to the infectious period are identical because the time of becoming infectious and time of recovery are assumed to be known in all cases.

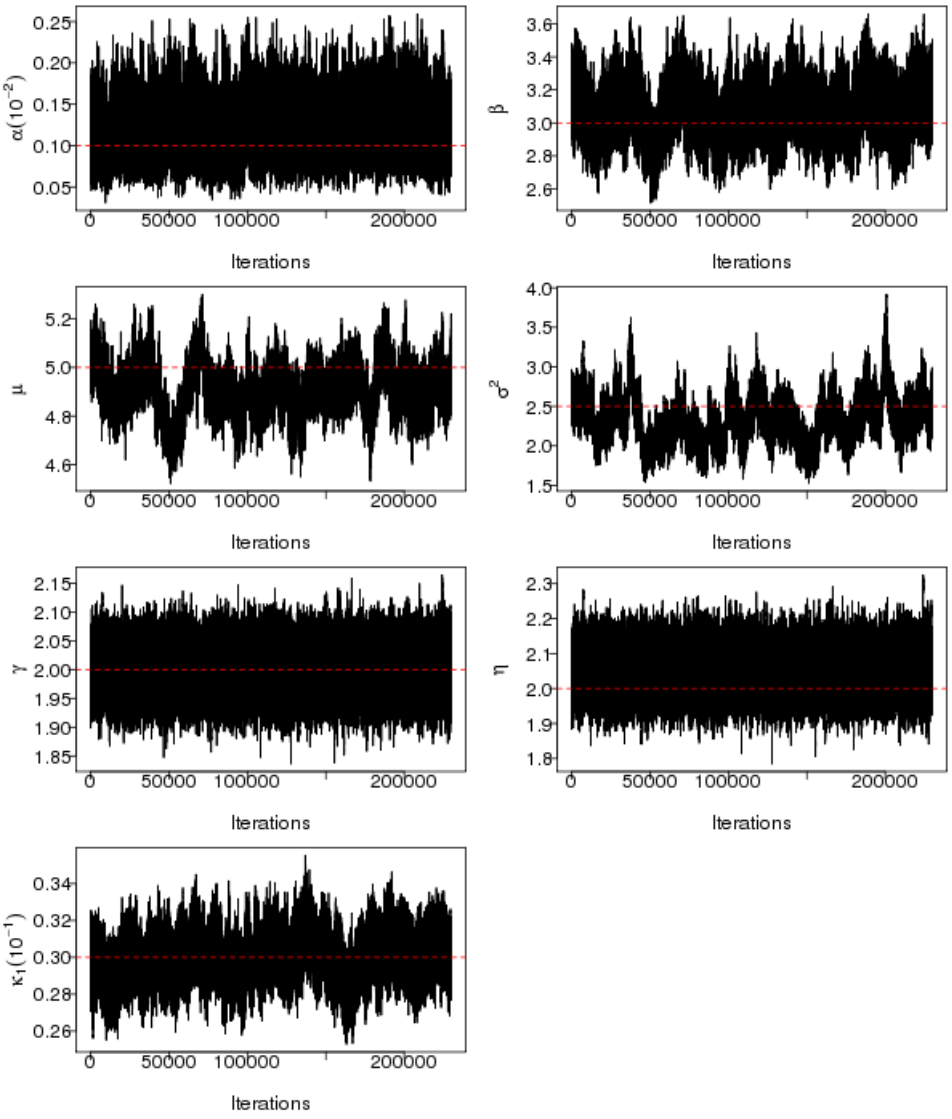


Figure 3.4: Traceplots of the posterior samples of model parameters obtained from fitting the correct model to the simulated data (Case A, Replicate 1). Dotted lines represent the actual values of the parameters used for simulating the epidemic.

Table 3.1: Values of $\pi(P(\widetilde{r}_j) < 0.05|y)$, $j = 1, 2, 3$, estimated from 1,500 posterior samples of the corresponding components of $\widetilde{\boldsymbol{r}}$ for 3 replicate epidemics simulated from the specified model and analysed using four different model assumptions (Case A, the correct model structure; Case B, a mis-specified (Cauchy-type) spatial kernel; Case C, a mis-specified (power-law) spatial kernel; Case D, a mis-specified latent period distribution).

	$\pi(P(\widetilde{r}_1) < 0.05\|y)$			$\pi(P(\widetilde{r}_2) < 0.05\|y)$			$\pi(P(\widetilde{r}_3) < 0.05\|y)$	
	Case B	*Case C*	*Case D*	*Case A*	*Case B*	*Case C*	*Case A*	*Case D*
Replicate 1	8%	5%	3%	6%	100%	80%	4%	99%
Replicate 2	5%	5%	4%	6%	100%	75%	4%	97%
Replicate 3	6%	6%	4%	5%	100%	76%	5%	100%

A *Weibull* infectious period is fitted in all cases. Uniform priors, which should be constrained to bounded regions to ensure a proper posterior distribution, are specified for all model parameters estimated in the following analyses.

In each case, we use standard MCMC and data augmentation to generate a sample from $\pi(\boldsymbol{\theta}, \boldsymbol{z}|y)$ from which - see section 3.2 - we impute posterior samples of the Infection-Link Residuals (ILR, $\widetilde{r}_2$) and of the Latent Time Residuals (LTR, $\widetilde{r}_3$). In addition we impute posterior samples of $\widetilde{r}_1$. The *Anderson-Darling* test for consistency with the uniform distribution is applied to each sample of the residuals. Posterior distributions of the model parameters are shown in detail in Figure 3.3. Figure 3.4 shows reasonable convergence and mixing on the basis of visual inspection.

Table 3.1 shows the values of $\pi(P(\widetilde{r}_j) < 0.05|y)$, $j = 1, 2, 3$, from three independently simulated replicates, y, of the epidemic. From $\pi(P(\widetilde{r}_2) < 0.05|y)$ and $\pi(P(\widetilde{r}_3) < 0.05|y)$ it appears that these posterior summaries systematically give evidence against the model when the spatial kernel and the latent period have been mis-specified respectively. On the other hand $\pi(P(\widetilde{r}_1) < 0.05|y)$ suggests no evidence against the model specifications in Cases B, C and D. Note that, for ease of comparison, we only present the value of $\pi(P(\widetilde{r}_j) < 0.05|y)$ for relevant cases. Values in cases not presented range from 3% to 6% and, therefore, suggest no evidence against the respective model specification.

3.4.1 Rationale for ordering the infection links

The ordering of infection links for an exposure event according to the strength of the links p_{ij} is not necessary when one only wishes to generate stochastic realisations of the epidemic using a functional-model representation. However the operation is crucial to ensuring that the imputed $\widetilde{r}_2$ may be informative regarding possible mis-specification of the kernel. In particular, our goal is to distinguish between the comparative goodness-of-fit of radially symmetric kernels that differ in terms of their behaviours at short and at long distances. When the ordering operation is used, the imputed residuals corresponding to infection links that represent very long or very short-range transmissions are located at the extremes of the unit interval. A mis-specified kernel, which misrepresents the propensity for short- or long-range transmission, may therefore be expected to cause the distribution of imputed $\widetilde{r}_2$ to deviate from $U(0, 1)$ by exhibiting a concentration, or a scarcity, of residuals at the extremes of the unit interval, dependent on the nature of the mis-specification. Our results presented in the next section suggest that this hoped-for sensitivity is indeed achieved.

To illustrate the point further we consider the case where an exponentially-bounded kernel is fitted to data generated using a power-law kernel (corresponding to Scenario I in the next section). Figure 3.5 shows a schematic representation of the relative strength of interaction of these kernels. The strength of infection links are represented by the segments length, and the monotonically decreasing nature of these kernels means that short range links are associated with stronger links than longer range interactions. Figure 3.5 shows the tendency of the exponential kernel to underestimate the interaction strength at short and long distances; as a result, the lengths of these infection links (which correspond to short and long distances transmission) are reduced when an exponential kernel is fitted. Should these links are imputed as the active links for exposure events, then the corresponding imputed residuals will be located closer to the extremes of the unit interval, than had the correct kernel been used, leading to a concentration of residuals at the extremes (also see Figure 3.6) in this case. Note that the above ordering operation is specifically aimed at comparing radially symmetric kernels with different tail properties, a common goal in epidemic studies. Alternative ordering schemes may be considered if our prior knowledge suggested a different form of mis-specification. For example, if transmission were to occur preferentially in certain directions, for example due to prevailing wind or other effects, then an ordering operation that took account of the direction, as well as the length, of the I-S links may be advisable. Such developments are beyond the scope of this thesis.

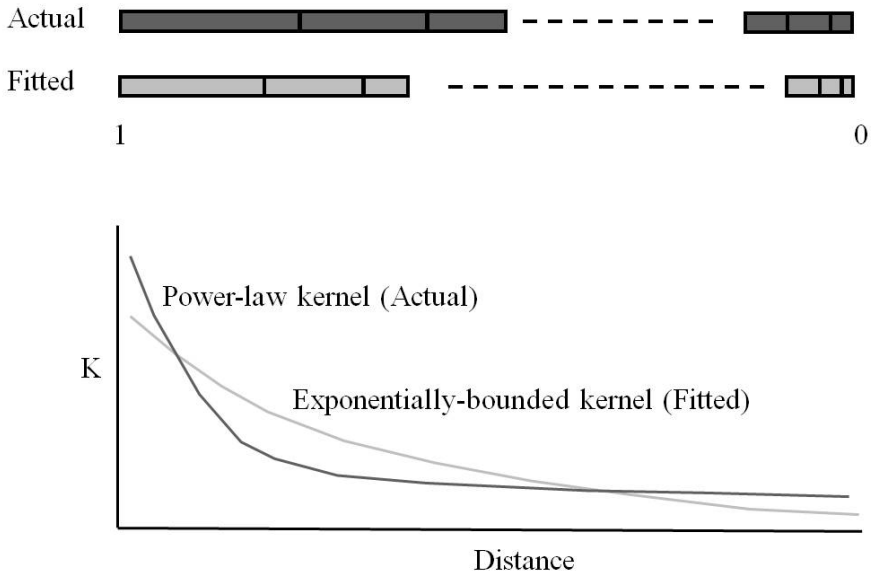


Figure 3.5: A schematic representation of the relative strength of interaction of these kernels. The strength of infection links are represented by the segments length.

3.4.2 Diagnosing model mis-specification

We now illustrate the insights our approach offers for understanding the causes of model inadequacy. Specifically, having observed considerable evidence against a model from the measure $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|y)$, we show that the pattern of residuals $\pi(\widetilde{r_2}|y)$ can suggest the manner in which the fitted model may be deficient. Consider two (symmetric) scenarios. In Scenario I, the epidemic is simulated from a kernel $K(d, \kappa_2) = d^{-2.8}$ and fitted with a kernel $K(d, \kappa_1) = \exp(-\kappa_1 d)$; in Scenario II, the epidemic is simulated from a kernel $K(d, \kappa_1) = \exp(-0.03d)$ and fitted with a kernel $K(d, \kappa_2) = d^{-k_2}$. Under the assumption that the fitted model is correct the imputed ILR should resemble samples from $U(0, 1)$. To highlight any systematic deviations from this null hypothesis Figure 3.6 presents the histogram formed by taking union of the subset of the ILR processes that produce small p-values (< 0.05) revealing a symmetry between the two scenarios. Scenario I and Scenario II respectively lead to a concentration or a scarcity of residuals at the extremes of the unit interval. This symmetry suggests that nature of the incompatibility of the spatial kernel may be diagnosed from systematically different deviations of the distribution of the ILR from $U(0, 1)$.

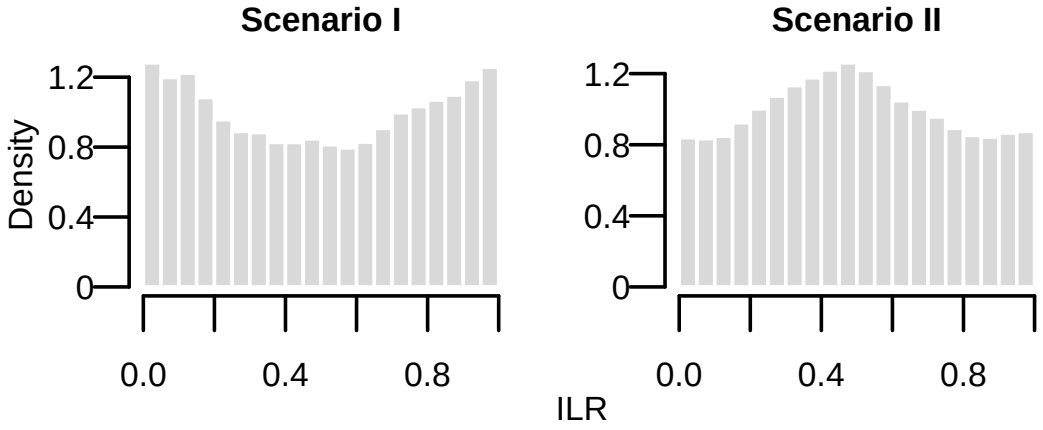


Figure 3.6: The distributions of a subset of imputed $\widetilde{r}_2$ whose $P(\widetilde{r}_2) < 0.05$ under two scenarios.

3.4.3 Comparison with common Bayesian model checking techniques

One common tool for model checking is the *Deviance Information Criterion* (*DIC*) (Spiegelhalter et al, 2002). The model with smallest DIC corresponds to the best model and, conventionally, models whose DIC exceeds that of the best model by

more than 10 units are considered to display substantial evidence of poor fit. A key limitation of this approach is that it is known to be problematic when applied to processes that are only partially observed, as in the case of most real-world systems, where the DIC cannot be uniquely defined (see 2.5 for a detailed discussion). Following (Celeux et al, 2006), we compute two versions of DIC,

$$DIC_1 = -4\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{X}}[log(f(\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}] + 2\mathbb{E}_{\boldsymbol{X}}[log(f(\boldsymbol{y},\boldsymbol{X}|\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}])|\boldsymbol{y}] \tag{3.21}$$

and

$$DIC_2 = -4\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{X}}[log(f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})|\boldsymbol{y}] + 2\mathbb{E}_{\boldsymbol{X}}[log(f(\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{X}))|\boldsymbol{y}] \tag{3.22}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ represent the unobserved and observed data respectively and $\hat{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{X})$ is often estimated by posterior point estimates such as the posterior mean which is used here (note that DIC_1 and DIC_2 are referred to as DIC_4 and DIC_8 in (Celeux et al, 2006)). The quantities $f(\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\theta})$ and $f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})$ represent contributions to the likelihood from both the observation model and the process model in the first case, and the observation model alone in the second. Notice that calculation of each version of the DIC requires expectations of these quantities which can be estimated using MCMC techniques. The main difference between DIC_1 and DIC_2 is that the first takes the unobserved data into account. However, there is no absolute theoretical justification for a preference of one definition over another.

Table 3.2 shows that, although both versions of DIC can differentiate the *relative* goodness-of-fit between Case A and Case C as well as that between Case A and Case D, DIC_2 misleadingly suggests that the fit for Case B is better than that for Case A. Note that the DIC is not a direct measure of model adequacy and only measures the relative goodness-of-fit between two models. Moreover, as shown in Table 3.2, the ranking of models can also vary between different versions of DIC.

Table 3.2: DIC computed for Cases A, B, C and D

Replicate 1	*Case A*	*Case B*	*Case C*	*Case D*
DIC_1	10357.52	11561.90	10542.75	11525.69
DIC_2	5754.594	4982.372	5937.58	6897.95

We further consider the performance of DIC and posterior predictive checks in an application to British floristic atlas data in Section 3.6.

3.5 Posterior predictive checking based on spatial autocorrelation analysis

In this section, we consider posterior predictive checks based on spatial autocorrelation coefficients which measure spatial dependency among observations, we specifically consider two common measures Moran's I and Geary's c indexes (Getis, 1991). The epidemic is simulated with kernel $K(d, \kappa) = d^{-2.8}$. We respectively fit a correct kernel $K(d, \kappa_2) = d^{-k_2}$ (Model 1) and an incorrect kernel $K(d, \kappa_1) = \exp(-\kappa_1 d)$ (Model II) to the simulated data. Predictive distributions of the spatial autocorrelation coefficients from Model I and Model II, at three different time points within the observation period are compared to the corresponding measures computed from the actual (simulated) data.

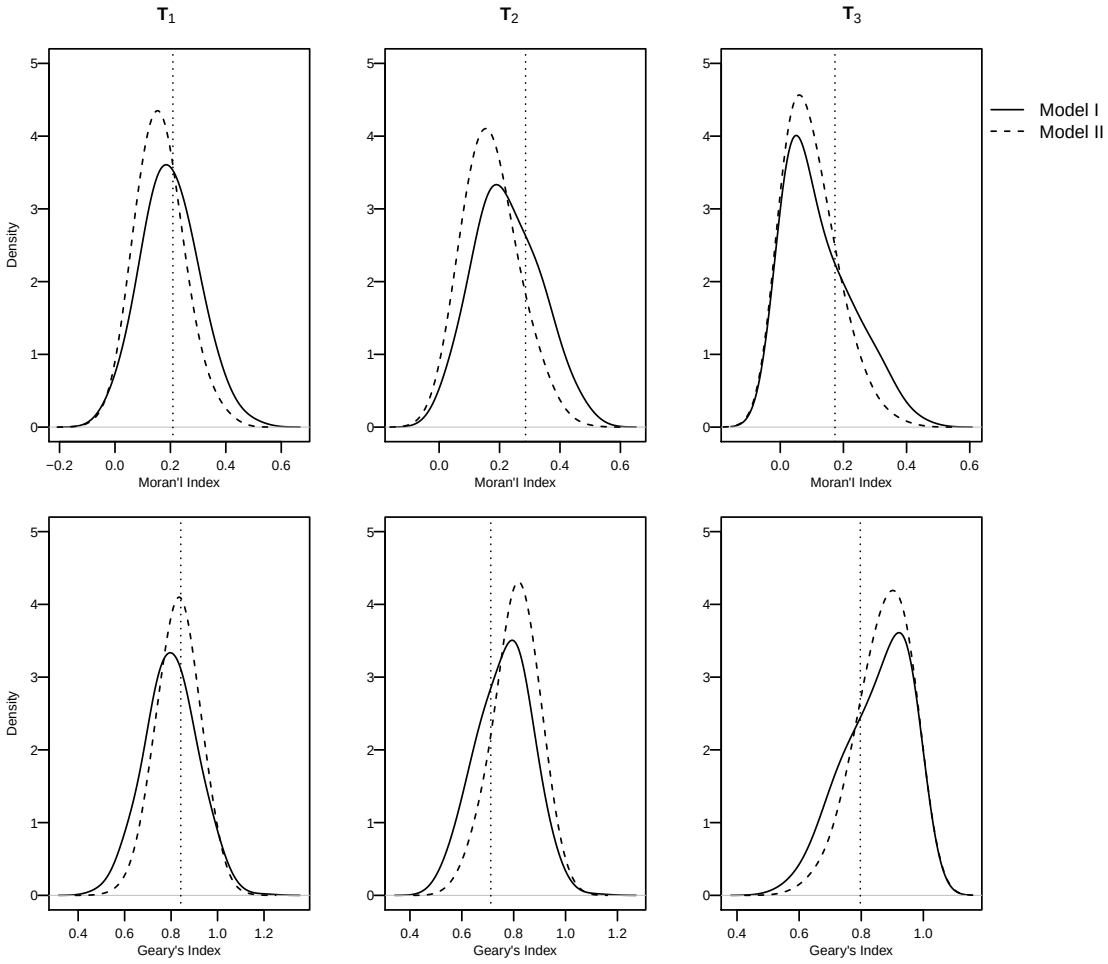


Figure 3.7: Posterior predictive distributions of Moran's I and Geary's c indexes obtained by simulating 1,000 epidemics from Model I and Model II respectively at time points $T_1 = 25$, $T_2 = 35$ and $T_3 = 45$. The vertical lines represent the observed values computed from the 'actual' epidemic.

We divide the 2000×2000 square area into $n = 100$ equally-sized square sub-regions

and count the number of sites x_i in class I or class R in sub-region i at time points considered for the computation of Moran's I and Geary's c indexes

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.23}$$

$$c = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.24}$$

where $\overline{x}$ is the mean of x_i over n sub-regions and w_{ij} is the spatial weight between sub-region i and j. There are many ways to define w_{ij}, and here we use the common binary weights in which $w_{ij} = 1$ if i is a *neighbour* of j, otherwise $w_{ij} = 0$. The definition of *neighbour* is also open to variations, and here we define that any sub-regions whose centroids are within two sub-region width (400) from the centroid of sub-region i are considered to be the neighbour of i. Both indices are computed by using a package *spdep* (Bivand et al, 2011) available in the statistical software R.

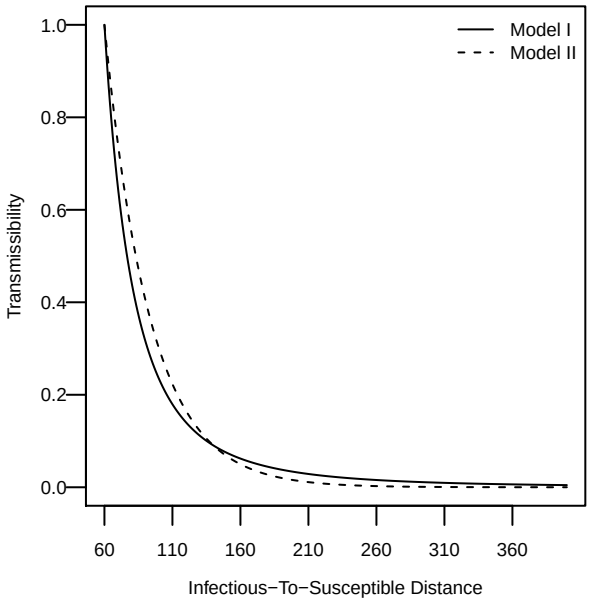


Figure 3.8: Estimated spatial kernels from fitting Model I and Model II with kernel parameters set to posterior means. Transmissibility is expressed relative to the amplitude of the respective kernel at $d = 60$ to highlight the difference between two kernels at both short and long distances.

Model I represents a long-tail dispersal mechanism and Model II represents a localised dispersal mechanism - this is also illustrated by Figure 3.8 - and $\pi(P(r_2) < 0.05|y)$ (i.e., the primary measure of degree of model mis-specification in utilizing r_2) shows strong evidence against Model II (90%) and no evidence against Model I. Figure 3.7 shows the predictive distributions of these two indices (at three different time points) obtained from simulating (1,000) epidemics respectively from Model I and Model II

with the model parameters drawn from their respective posterior distributions. It can be seen from the figure that the posterior predictive distributions of the spatial autocorrelation indices from both models are broadly consistent with the observed values (i.e., all of the 95% two-sided intervals contain the actual value). This shows that posterior predictive checks based on these indices when only partially observed epidemics are available could be insensitive to the specification of the spatial kernel. Posterior predictive checking has to be computed 'offline' - for example, one needs to obtain the posterior distribution or point estimates of model parameters first and compute the required summary statistics based on simulation techniques. Our method, instead, can be easily embedded along with the estimation and by default takes the full posterior distribution of model parameters into account. It is also noted from above that summary statistics based on these spatial autocorrelation measures are subject to variations in definitions which might lead to different conclusions.

3.6 Case Study: Spread of Giant Hogweed in Great Britain

Invasive alien species represent a major threat to ecosystems and cause significant environmental and financial loss worldwide (Pimentel et al, 2005; Vilà et al, 2009). *Heracleum mantegazzianum* (giant hogweed) causes significant problems in Great Britain and has rapidly spread since 1970 (Catterall et al, 2012). We apply our testing framework to British floristic atlas data which assess the presence of giant hogweed over a square lattice of 10×10 km resolution in 1970, 1987 and 2000. In total 2,838 such squares are considered to be habitable for the giant hogweed (see (Catterall et al, 2012)). These are classified as susceptible or colonised at the given survey times according to the absence or presence of giant hogweed in the lattice. These data are well suited to testing our methodology. Detection of the species is relatively easy due to its height ($>$ 2m), so that the number of false absences in the data set should be limited. Moreover, the data give 'snapshots' of the distribution at three distinct times (from 1970 to 2000) and over a large region making them particularly suitable for inferring the spatio-temporal transmission mechanism.

We first represent the lattice of square regions as a lattice of points where the position of a point is given by centre of the square which it represents. Figure 3.9 shows the snapshots of the spread of giant hogweed in Great Britain taken at three distinct times (1970, 1987 and 2000). In the light of the aggressive nature of giant hogweed we assume that, once colonised, sites can immediately start to colonise other sites and remain colonised. In the terminology of the epidemic model we consider, there-

Figure 3.9: Snapshots of the spread of giant hogweed in Great Britain taken at three distinct times: (a) 1970, (b) 1987 and (c) 2000. Black dots represent the colonised sites.

fore, the E and I classes to be a single class and dispense with the recovery class R from our model. Effectively, we fit an S-I (Susceptible-Infectious) model to the presence/absence data and use our model assessment methods to compare the goodness-of-fit of several formulations, discussed in detail in Section 3.6.3. In summary, the models differ in the choice of spatial kernel and with regard to the inclusion of terms quantifying the suitability of each site, j, for the species. Suitability is represented by a measure $c_j \in [0, 1]$, where the c_j are taken from an earlier analysis (Catterall et al, 2012) in which an extensive range of covariates including average temperature, altitude and other factors, were considered in their estimation. With suitability included, the instantaneous rate at which a susceptible site j becomes colonised Equation (3.1) is moderated by a factor c_j.

Full model specifications and posterior estimates of the model parameters are described in Section 3.6.3. Specifically, we consider three forms of spatial kernel with and without homogeneous suitabilities giving rise to six models:

- Model 1 (M1, Kernel A): $K(d_{ij}, \kappa_1) = exp(-\kappa_1 d_{ij})$, heterogeneous suitabilities, c_j;
- Model 2 (M2, Kernel B): $K(d_{ij}, \kappa_2) = 1/(1+d_{ij}/\kappa_2)$, heterogeneous suitabilities, c_j;
- Model 3 (M3, Kernel C): $K(d_{ij}, \kappa_3) = d_{ij}^{-\kappa_3}$, heterogeneous suitabilities, c_j;
- Model 4 (M4, Kernel A): $K(d_{ij}, \kappa_1) = exp(-\kappa_1 d_{ij})$, homogeneous suitabilities

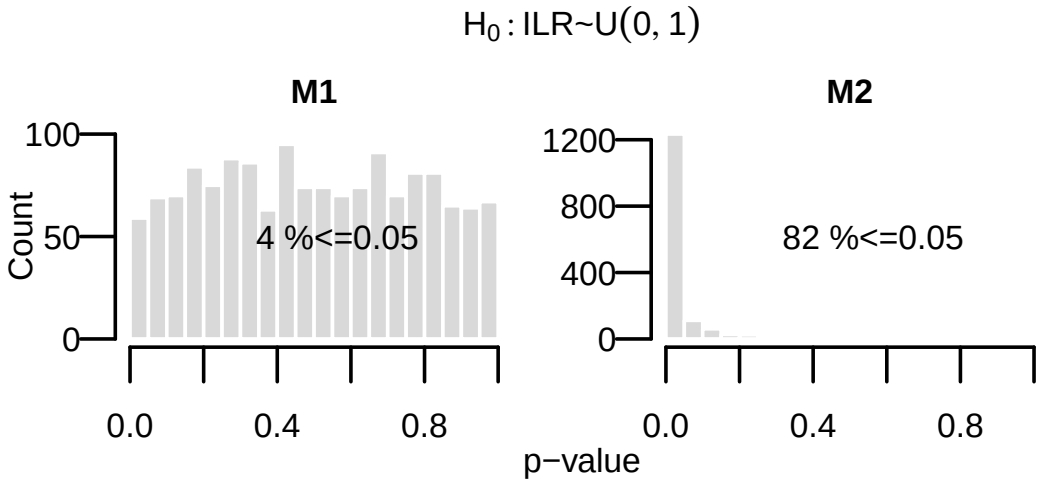


Figure 3.10: Posterior distributions of the p-values from testing the sets of posterior samples of Infection Link Residual (ILR) imputed from MCMC chains (1,500 samples in each case) when fitting SI models, representing heterogeneous suitability, to the giant hogweed data with kernel A (model M1) and kernel B (model M2) respectively.

$c_j = 1$;

- Model 5 (M5, Kernel B): $K(d_{ij}, \kappa_2) = 1/(1 + d_{ij}/\kappa_2)$, homogeneous suitabilities $c_j = 1$;
- Model 6 (M6, Kernel C): $K(d_{ij}, \kappa_3) = d_{ij}^{-\kappa_3}$, homogeneous suitabilities $c_j = 1$.

These models are fitted to the data using Bayesian methods as described above. For the simple SI formulation the residual process reduces to $\widetilde{\boldsymbol{r}} = (\widetilde{r}_1, \widetilde{r}_2)$ and specifies exposure times and infection links. We apply three tests to imputed values of these residuals for each of the models. As with the simulated example, we investigate $\pi(P(\widetilde{r}_1) < 0.05|y)$ and $\pi(P(\widetilde{r}_2) < 0.05|y)$ arising from an Anderson-Darling test applied to the respective subset of residuals. We also consider a combined test, with p-value $P(\widetilde{\boldsymbol{r}})$, based on a test statistic

$$T(\widetilde{\boldsymbol{r}}) = -2\left(\log(P(\widetilde{r}_1)) + \log(P(\widetilde{r}_2))\right)$$

whose distribution under the modelling assumptions is χ^2_4, and report $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|y)$ for each model. Conclusions arising from the various tests are presented in the following subsections.

Table 3.3: Values of $\pi(P(\widetilde{r}_j) < 0.05|y)$, $j = 1, 2$, and $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|y)$ estimated from 1,500 posterior samples of ILR and ETR computed from the giant hogweed data under different model assumptions regarding the spatial kernel and dependence on suitability of sites

	$\pi(P(\widetilde{r}_1) < 0.05\|y)$			$\pi(P(\widetilde{r}_2) < 0.05\|y)$			$\pi(P(\widetilde{\boldsymbol{r}}) < 0.05\|y)$		
Spatial Kernels	A	B	C	A	B	C	A	B	C
Heterogeneous Suitability	13%	7%	11%	4%	82%	4%	11%	74%	10%
Homogeneous Suitability	5%	6%	3%	35%	100%	54%	24%	100%	28%

3.6.1 Model assessment and implications for control strategies

From Table 3.3 we first notice that, regardless of whether dependence on suitability is included, $\pi(P(\widetilde{r}_2) < 0.05|y)$ and $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|y)$ is larger for the models with Cauchy-form kernel (Kernel B, M2 & M5) than for the respective models with exponentially bounded kernel (Kernel A, M1 & M4) or with power-law kernel (Kernel C, M3 & M6), suggesting that the Cauchy kernel typically provides a poorer fit. When dependence on suitability is not included (M4, M5 & M6), the fact $\pi(P(\widetilde{r}_2) < 0.05|y) = 0.35$ for M4 (Kernel A) and $\pi(P(\widetilde{r}_2) < 0.05|y) = 0.54$ for M6 (Kernel C) suggests there are substantial probabilities that the U(0,1) hypothesis for the imputed residuals would be rejected by the classical observer and calls these models into question. By contrast, the results for M1, M2 and M3 present no evidence against the model with exponentially bounded kernel (A) and power-law kernel (C), while there remains a substantial posterior probability of rejection (0.82) for the model with Cauchy-form kernel (B). Figure 3.10 presents samples from $\pi(P(\widetilde{r}_2)|y)$ for M1 and M2, highlighting the evidence against kernel B. It is clear from other results in Table 3.3 that the evidence against a given model arises from the ILR residuals $\widetilde{r}_2$; the test based on $\widetilde{r}_1$ alone presents little evidence against any of the models M1-M6, while the evidence arising from the combined test is typically weaker than that from the tests of $\widetilde{r}_2$ alone.

In summary we find evidence that the dispersion mechanism for hogweed cannot be represented adequately by the Cauchy dispersal kernel, while no evidence against the exponentially-bounded kernel and power-law kernel is found as long as habitat heterogeneity is accommodated. Figure 3.12 shows that the long-tail behaviour of M2 tends to induce a scarcity of residuals at the left end and a concentration of residuals at the right end of the unit interval. Although the ILR residuals $\widetilde{r}_2$ were constructed with assessment of spatial kernels in mind, comparison of $\pi(P(\widetilde{r}_2) < 0.05|y)$ between models with and without the dependence on suitability (i.e., comparison between M1 and M4 and that between M3 and M6) highlights the potential for the method to detect mis-specification of other aspects of the model. This is not surprising given the key role of $\widetilde{r}_2$ and the suitabilities in the construction of the colonisation links.

Giant hogweed spread their seed mostly through wind, water and human activities (Pysek et al, 2007). Localised dispersal mechanisms typically involve the dispersal by wind or animal activities. Human activities, such as soil transport and transport of seeds adhering to the car tyres, are mainly responsible for long-distance dispersal. Understanding of the importance of short-distance and long-distance dispersal provides valuable insight for devising appropriate control strategies (Dawe and White,

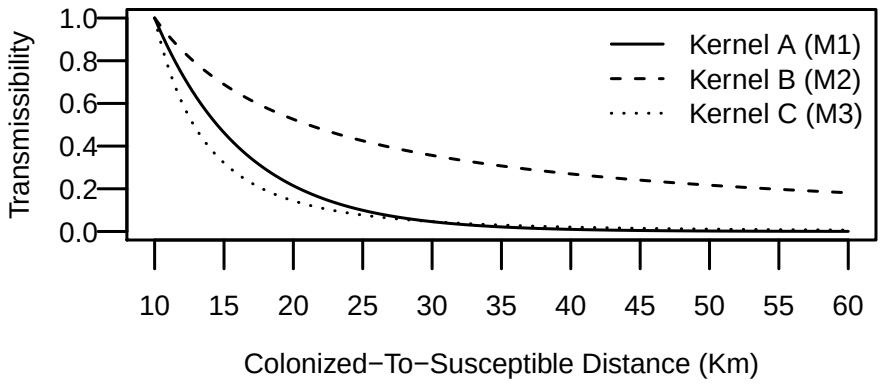


Figure 3.11: Estimated spatial kernels from fitting M1 and M2 and M3 to the giant hogweed data with kernel parameters set to posterior means. Transmissibility is expressed relative to the amplitude of the respective kernel at 10km (the minimum distance between two sites in the data set)

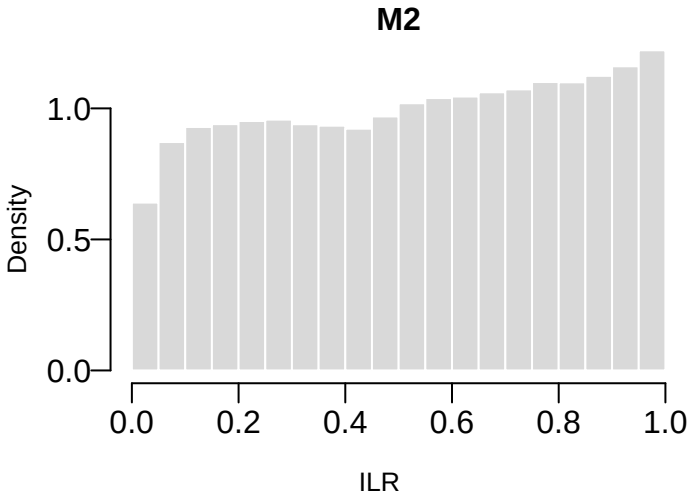


Figure 3.12: Distributions of subsets of imputed ILR which lead to p-values less than 0.05 from M2.

1979; Pergl et al, 2011). Our results and Figure 3.11 clearly suggest that the spread of hogweed is mainly via a nearest-neighbour mechanism. Given this highly localised dispersal mechanism, and the constraints imposed by the lattice structure of the hogweed data (Section 3.6) which forces a minimum distance of 10km between two sites (in contrast to a more general continuous space in the simulated example), the difference between an exponentially-bounded kernel and a power-law kernel becomes insignificant (see Figure 3.11). Hence the two models display similar goodness-of-fit. This suggests that control measures - for instance, education programs of increasing public awareness and participation in prevention and reporting (Bhowmik, 2005), and field survey and subsequent eradication measures (Sampson et al, 1994) - may be

most effectively deployed by focusing implementation on the neighbourhood of known colonisations.

3.6.2 Comparison with DIC and posterior predictive checks

Similar to Section 3.4, we compute two versions of DIC, DIC_1 and DIC_2 (see Equation (3.21) and Equation (3.22)), for M1 and M2 and the corresponding models M4 and M5 which do not take the suitability into account. From Table 3.4, we note that, although both versions of DIC can differentiate the *relative* goodness-of-fit between M1 and M2, they unreasonably indicate that M4 (exponentially-bounded kernel without considering suitability) is the best or the second best model.

In Section 3.4, we have shown that summary statistics quantifying spatial autocorrelation may be less sensitive to model mis-specification than our latent residual approach. We therefore focus on more intuitive summary statistics based on the number of colonisations. We also adopt a conservative approach, running forward simulations of the fitted model using point estimates of parameters, here the posterior mean. Moreover, these simulations are conditioned on the colonised sites observed in 1970. We focus on comparing models M1 and M2 (which include dependence on suitability and for which our residual analysis shows a significant difference in goodness-of-fit) and examine the following predictive outcomes: the predictive distribution of the number of colonised sites, at the end of the observation period (2000), within annular regions centered on a given location (see Figure 3.13 for a representation of the regions), and the numbers of reported colonised sites at the second and third observation times (1987 and 2000). Figure 3.14 and Table 3.5 compare predicted distributions and the actual observations.

Table 3.4: DIC computed for M1, M2, M4 and M5.

Model	M1	M2	M4	M5
DIC_1	7404.8	7442.9	7422.0	7890.1
DIC_2	968.3	1027.4	522.2	1194.8

Table 3.5: Predicted and reported new colonised sites at second and third snapshots (1987 and 2000). The reported numbers are followed by the two-sided 95% credible intervals enclosed in brackets (1,000 simulations for each model)

Model	M1	M2
1987	334 (311, 375)	334 (310, 381)
2000	412 (368, 434)	412 (388, 460)

Figure 3.13: The partition of Great Britain according to intersection with 65 concentric annuli. Each annulus is centred on the black dot and has width 10km

Figure 3.14 and Table 3.5 suggest that, similar to Section 3.4.3, model checks based on these apparently reasonable summary statistics may be insensitive to the choice of model. Figure 3.11 shows that M1 and M2 represent very different transmission mechanisms, with kernel B exhibiting a strong propensity for long-range transmission. Figure 3.15 shows that the estimates of transmission rates can be different when different kernels are fitted. Nevertheless, the predictive distributions of the summary statistic appear consistent with observed values for both M1 and M2.

3.6.3 Details of model formulation and posterior distributions of model parameters

The second snapshot at 1987 is known to be incomplete due to the insufficient efforts of surveying the sites. We therefore also estimate the probability, p, that a site has been colonised by 1987 but remained unreported given that it is reported at the last snapshot (2000).

Figure 3.16 shows the posterior distributions of the model parameters from the "best"

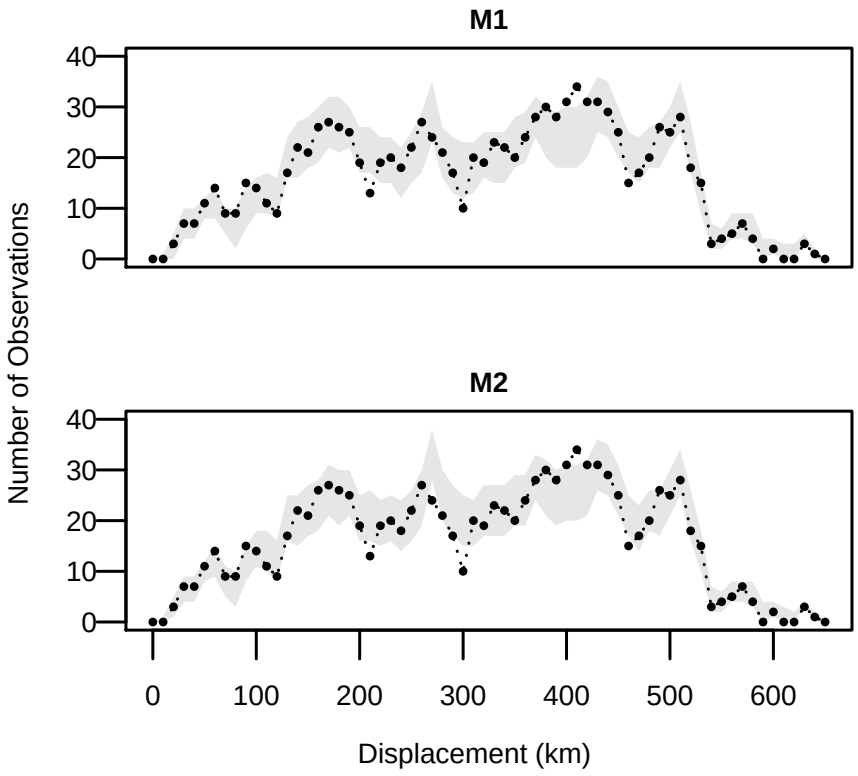


Figure 3.14: Distribution of the number of colonised sites within each ring region at the final observation time as predicted by models (the shaded area represents the 95% two-sided interval of the predicted number of colonised sites from 1,000 simulations) and the observed data (shown in the dotted black line). The displacements are measured from the center of the ring regions

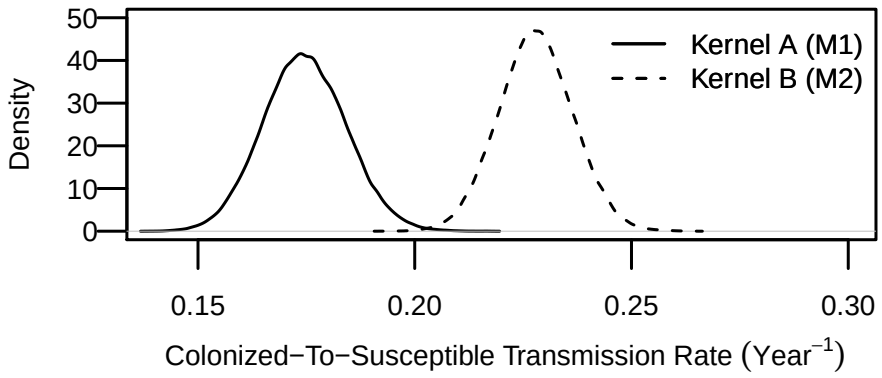


Figure 3.15: Posterior distributions of transmission rates from a colonised site to a susceptible site from fitting M1 and M2 to the giant hogweed data

and "worst" models − M1 (kernel A) and model M2 (kernel B) with the consideration of suitability. Corresponding Traceplots are shown in Figure 3.17 and Figure 3.18. Geweke's convergence diagnostic (Geweke et al, 1991) is also applied to the posterior samples of all model parameters (taking the first 10% and last 50% of the the chains),

and we obtain Z-score indicating convergence.

To reinforce the confidence over the specification of the suitability, we subdivide the sites into three classes (assuming common suitability in each class) according to the estimated suitability from the analysis (Catterall et al, 2012) and estimate the common suitability in each class.

Sites are classified into three classes (Less Favorable, Favorable and Highly Favorable, with suitability s_1, s_2 and s_3 respectively) according to the corresponding suitability estimated from the earlier study. We denote c_j as the estimate of suitability of site j given in Catterall et al (2012). If $0 < c_j \leq 0.25$, the site is classified as Less Favorable; if $0.25 < c_j \leq 0.5$, it is classified as Favorable; if $0.5 < c_j \leq 1.0$, it is classified as Highly Favorable. We estimate s_1 and s_2 as the suitability relative to $s_3 = 1$.

We consider fitting a model with kernel A (a 'better' kernel as we have shown). Figure 3.19 shows the posterior distributions of parameters s_1 and s_2. From the figure, it is evident that $s_1 < s_2 < s_3 = 1$ and it indicates that the previous estimates of suitability (Catterall et al, 2012) are reliable and broadly consistent with our estimates, which reinforces the confidence of adopting these earlier estimates (Catterall et al, 2012) for our model.

3.7 Limitations and extensions

We have so far presented a novel model assessment framework by assessing the posterior distributions of the latent residuals. In this section, we first discuss and investigate a confounding issue of this approach. To be specific, we investigate how the residuals testing performs when one wishes to test more than one model component at the same time. We also extend the testing framework to develop a sequential procedure of the latent-residuals test. Results show that this sequential procedure, in contrast to the non-sequential approach which always uses the full sample, may exhibit higher sensitivity in scenarios when the observations in the early stage of the epidemic encapsulate more information of the transmission dynamics. Particularly, we show that the Exposure Time Residuals (ETR), which do not appear to be sensitive in the full-sample approach, may become sensitive under this procedure.

3.7.1 A confounding issue

Although each element in $\widetilde{\boldsymbol{r}} = (\widetilde{r}_1, \widetilde{r}_2, \widetilde{r}_3, \widetilde{r}_4)$ is constructed specifically to assess a particular aspect of the model, there are potential confounding effects interplaying among

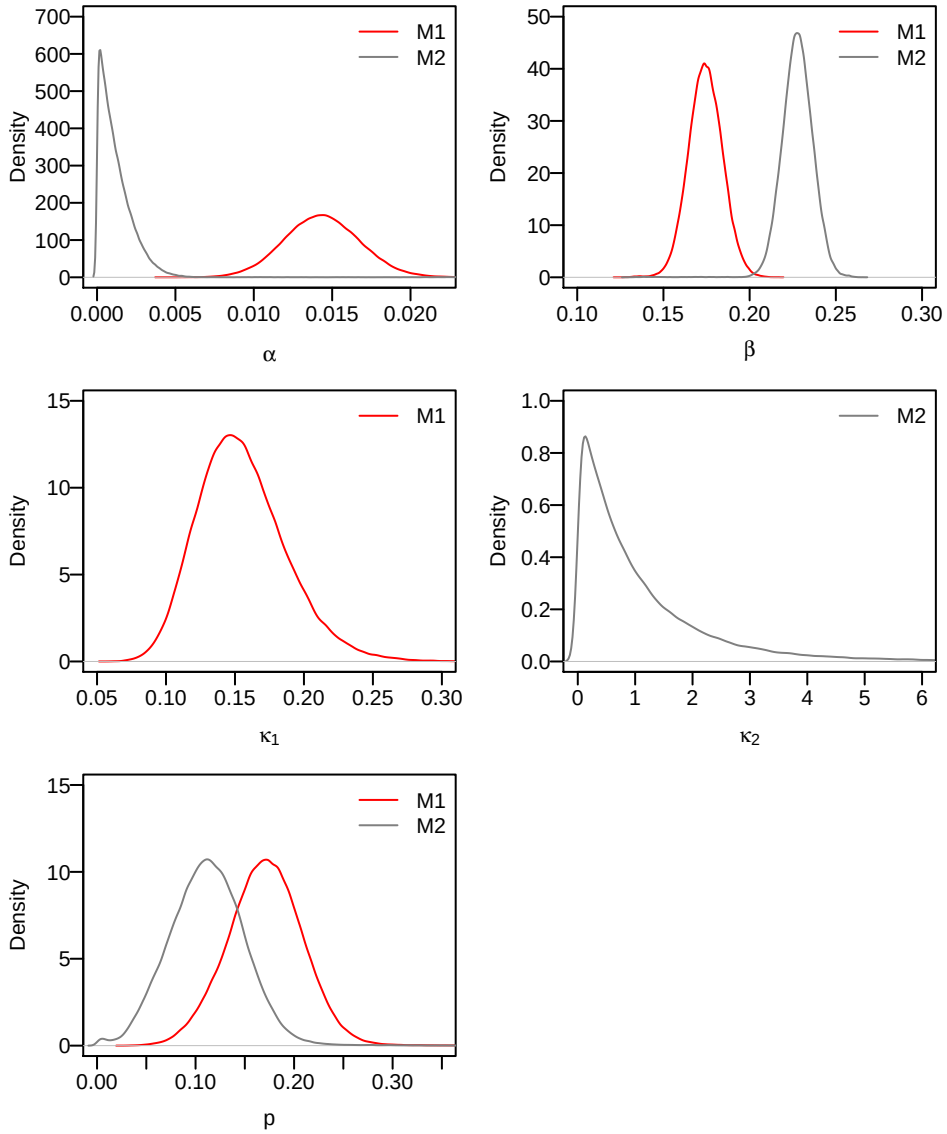


Figure 3.16: Posterior distributions of model parameters for models fitted to giant hogweed data where suitability of sites are considered.

these residuals. In practice, the "correct" structure of other model components (not being tested) may not be always known but may have to be assumed or estimated. The reliance on the assumption of other model components and the estimation of the corresponding model parameters gives rise to a potential confounding issue. For example, in the context of MCMC, the imputation of ILRs depends on other model assumptions such as infection times and the latent period distribution which might need to be estimated, and these estimates also depend on the spatial component specified − as a result, these model components can act as free parameters which allow a mis-specified spatial kernel to tune itself to fit the data better, hence exhibiting a better fit of the imputed ILRs with the uniform distribution and suggesting less evidence of mis-specification, compared to the situation where the correct structure of other model components are adopted. The effects due to this kind of interaction between a

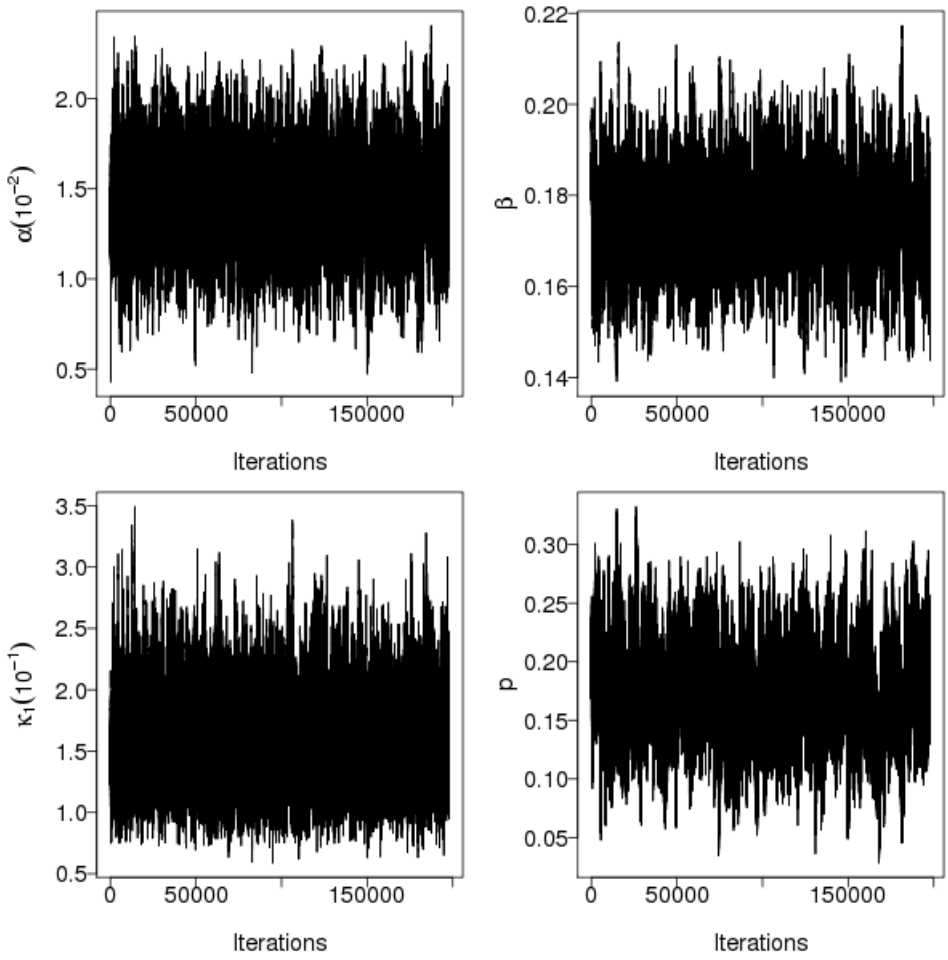


Figure 3.17: Traceplots of the posterior samples of model parameters obtained from fitting model M1 to giant hogweed data where suitability of sites are considered (with burn-in length 10,000).

particular type of residual and other aspects of the model are briefly explored in this section.

To investigate the confounding effects, we simulate an epidemic from the model specified in section 3.4 and consider four cases: in Case $CK\&CL$, a correct spatial kernel (Exponentially-bounded) and a correct latent period distribution (Gamma) are fitted to the data; in Case $WK\&CL$, a wrong spatial kernel (Cauchy) and a correct latent period distribution are fitted; in Case $CK\&WL$, a correct spatial kernel and a wrong latent period distribution (Exp) are fitted; finally in Case 4 $WK\&WL$, the wrong spatial kernel and the wrong latent period distribution are fitted.

The confounding effect of the specification of a latent period distribution on the ILR can be discerned from Figure 3.20. We notice that the sensitivity of the test based on ILR tends to decrease when a wrong latent period is specified, although a strong evidence against mis-specification is still observed. A similar confounding effect of the specification of a spatial kernel on the LTR is found (Figure 3.21). The reduction

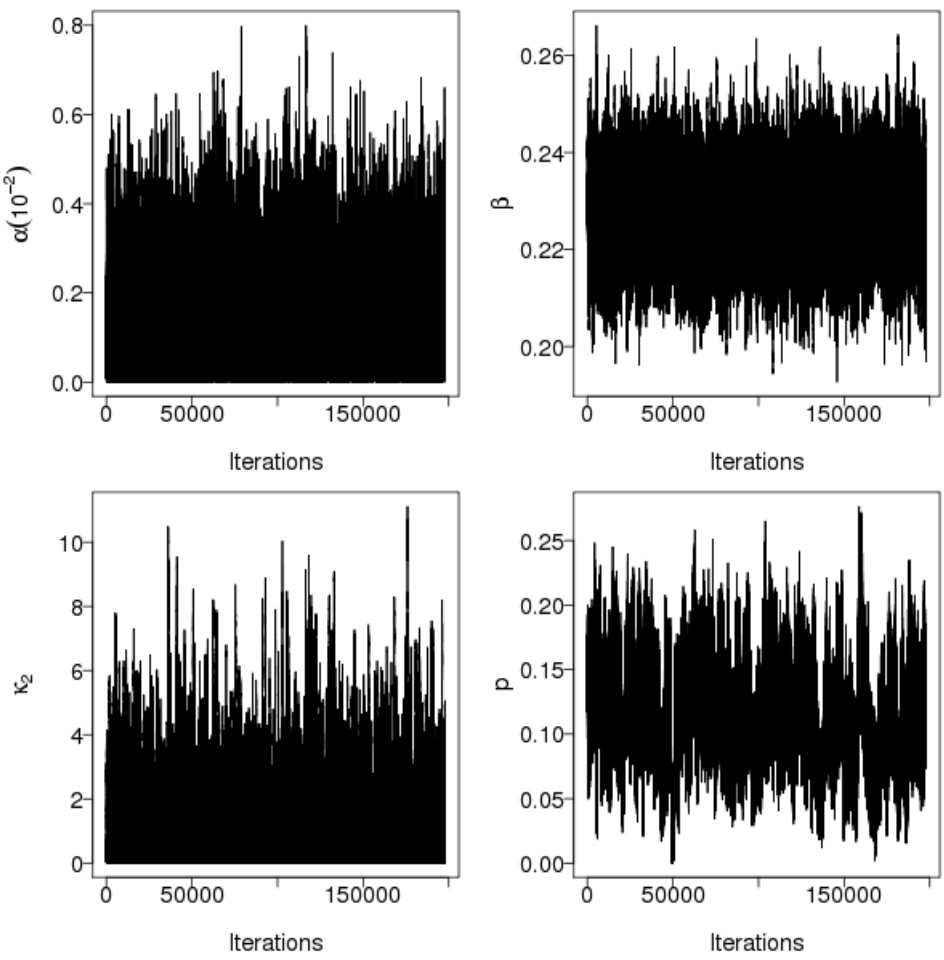


Figure 3.18: Traceplots of the posterior samples of model parameters obtained from fitting model M2 to giant hogweed data where suitability of sites are considered (with burn-in length 10,000).

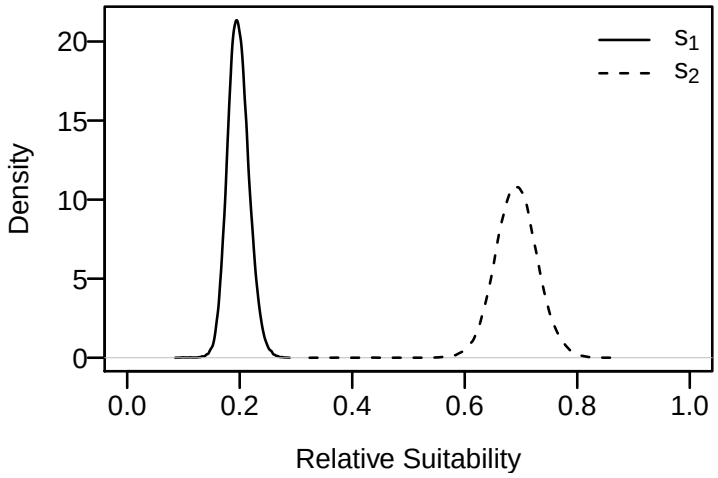


Figure 3.19: Posterior distributions of suitability parameters in the model (with kernel A) fitted to the giant hogweed data in which sites are classified into three classes

of evidence appears to be more significant in the test based on LTR than that based on ILR.

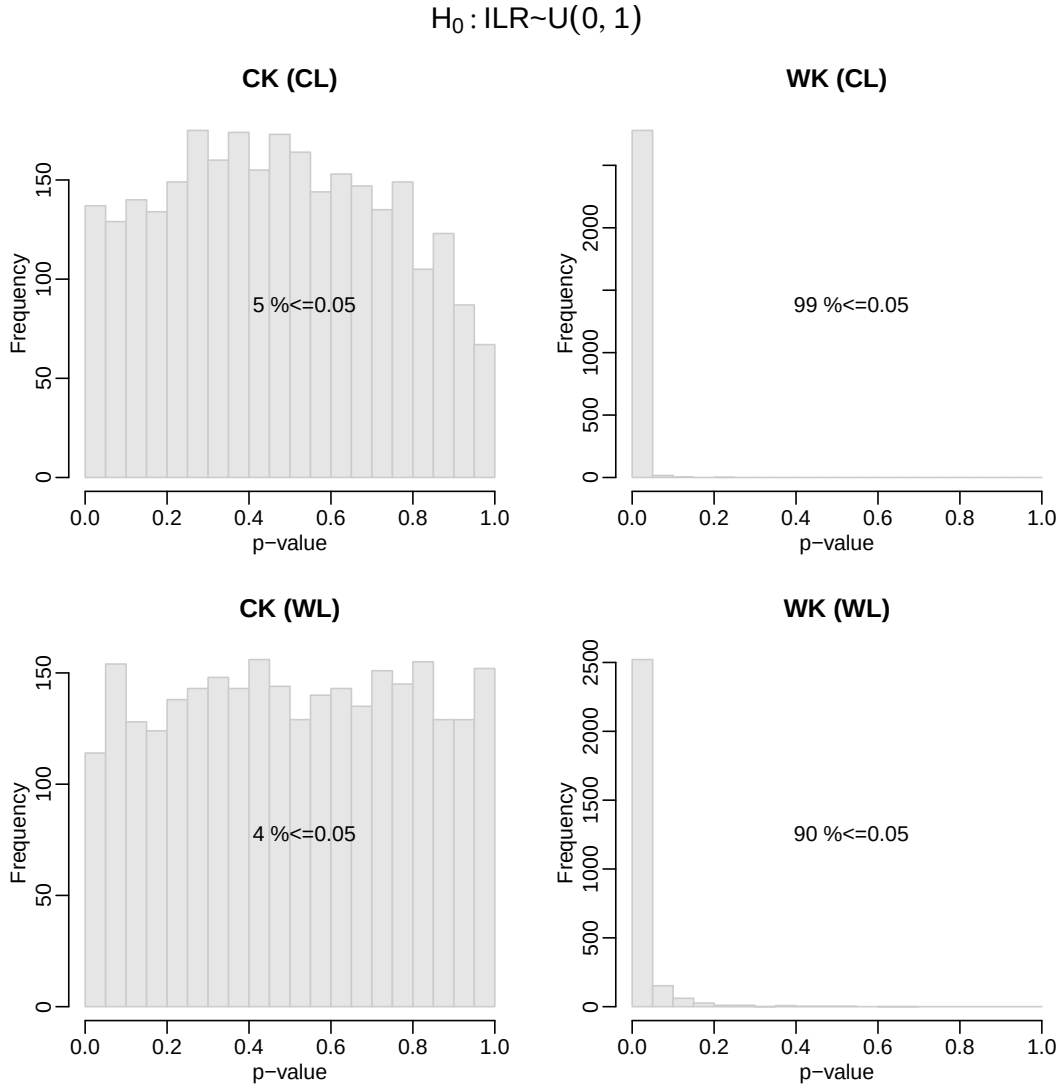


Figure 3.20: Posterior distributions of the p-values from testing the sets of posterior samples of Infection Link Residual (ILR) imputed from MCMC chains in fitting different combinations of the spatial kernel and the latent period.

3.7.2 Sequential latent-residual testing

The motivation behind the development in this section is two-fold. The latent-residual testing we have presented so far takes into account the full posterior residual sample whose *size* (i.e., size of the epidemic) is actually a random realisation of the epidemic process. From the point of view of a classical observer, it is, rigorously speaking, not appropriate to claim that the testing has a *known* type I error α that is defined in a hypothetical *repeating sampling* with a fixed sample size. To eliminate the effect of random sample size on the type I error, an alternative approach will be to consider

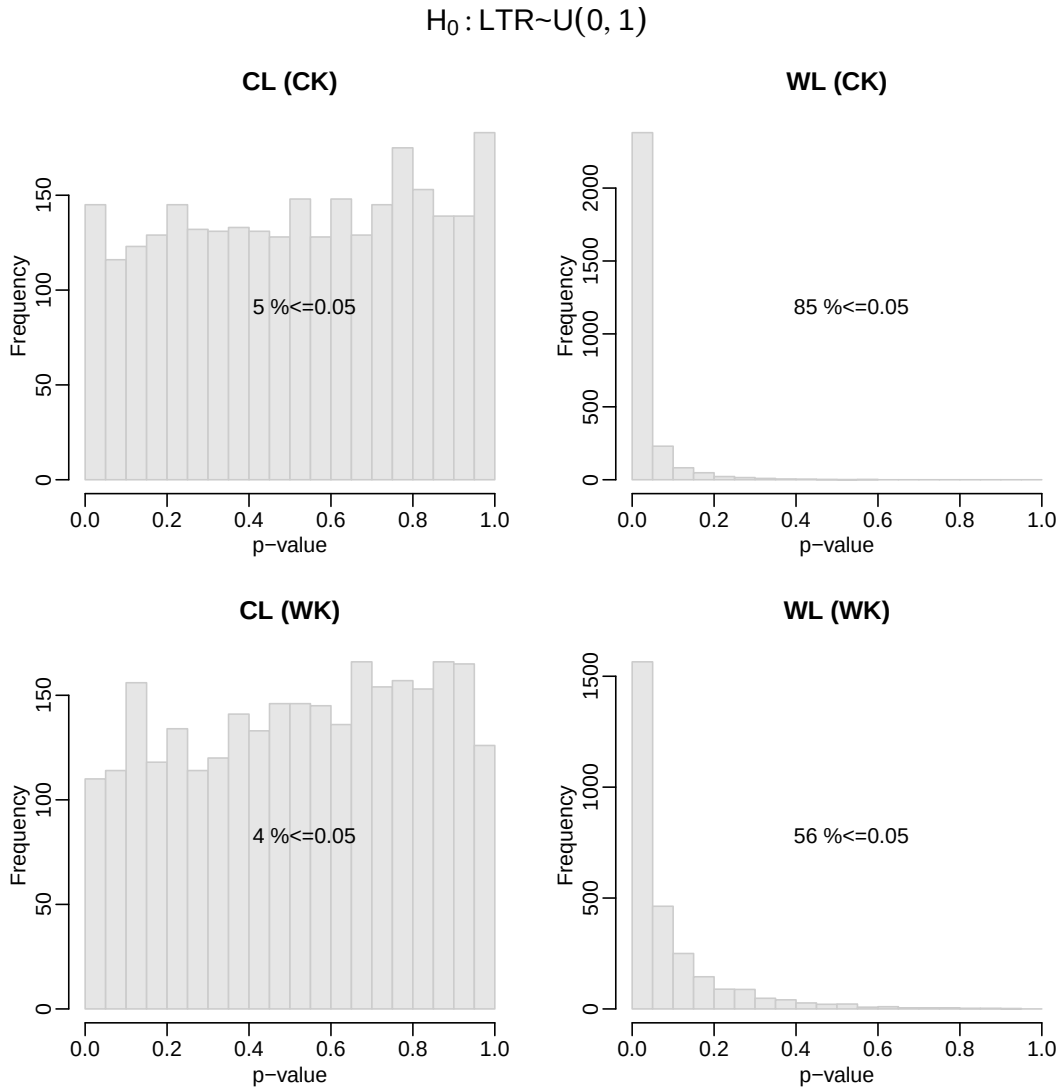


Figure 3.21: Posterior distributions of the p-values from testing the sets of posterior samples of Latent Time Residual (LTR) imputed from MCMC chains in fitting different combinations of the spatial kernel and the latent period.

sequential testing (aiming at a desired type I error rate) on a stepwise increasing subset of the full sample, where a decision may be made (by a classical observer) before the testing procedure reaching the end of the full sample. While this first part of the motivation is rather subtle and mainly of academic interest, its second part is mainly driven by a practical consideration − i.e. one may ask that whether the exposed cases during the early stage of an epidemic outbreak encapsulate more information of the transmission dynamics and hence lead to a more sensitive test against a wrong model. Or in other words, would the full-sample non-sequential approach potentially dilute the evidence of model mis-specification (if any) presented at the early stage? A sequential testing procedure on the residual sample sorted in ascending order of corresponding exposure times provides a natural framework for answering this question − this is also the primary problem we would like to investigate

in this section.

Evidence of model mis-specification in the early stage of an epidemic outbreak

Here we propose a sequential approach which may be more sensitive, when there is more evidence of model mis-specification contained in the early observations of an epidemic, than the non-sequential approach presented earlier in the chapter.

Consider m subsets $A_1 \subset A_2 \cdots \subset A_m$ of the posterior sample of the residuals $\widetilde{\boldsymbol{r}}$ sorted in ascending order of corresponding exposure times. Apply sequentially the Anderson-Darling test for $H_0 : A_i \sim U(0,1)$ where $i = 1, 2, \ldots, m$, with the testing procedure ceasing either when the Anderson-Darling statistic AD_i at step i exceeds a corresponding critical value c_i or when $i > m$. Recall that we use $\pi(P(\widetilde{\boldsymbol{r}}) \leq 0.05|y)$ to represent the posterior belief of a Bayesian observer regarding the p-value $P(\widetilde{\boldsymbol{r}})$ that a classical observer of $\widetilde{\boldsymbol{r}}$ would compute. Should this distribution be concentrated on small values, the Bayesian observer would infer that the classical observer may reject the hypothesis that the $\widetilde{\boldsymbol{r}}$ were generated as a random sample from a $U(0,1)$ distribution. When using the sequential approach proposed above, we look at the rejection frequency of the classical observer $\pi(T(\widetilde{\boldsymbol{r}}) = 1|y)$ where

$$T(\widetilde{\boldsymbol{r}}) = \begin{cases} 1, & \text{if } AD_k \geq c_k \text{ for some } k \leq m, \\ 0, & \text{if } AD_i < c_i \text{ for all } i \leq m. \end{cases} \tag{3.25}$$

A simulation-based approach

The key challenge of this approach is to specify appropriate c_i such that the desired type I error α, say, 5%, is achieved. We consider an approach in which we have a constant "local" type I error α' at each step of the sequential test. Specifically, we can solve the equation for α'

$$\begin{aligned} \alpha' + (1-\alpha')\alpha' + \cdots + (1-\alpha')^{m-1}\alpha' &= \alpha \\ 1 - (1-\alpha')^m &= \alpha, \end{aligned} \tag{3.26}$$

where the i^{th} term on the left hand side (of the first line) represents the probability that the null hypothesis is rejected at the i^{th} step but not at preceding steps. Subsequently,

c_i for $i = 1, 2, \ldots, m$ may be derived from the following equations

$$\begin{aligned} &P(AD_1 \geq c_1) = \alpha' \\ &P(AD_2 \geq c_2 | AD_1 < c_1) = \alpha' \\ &\vdots \\ &P(AD_m \geq c_m | AD_1 < c_1, \ldots, AD_{m-1} < c_{m-1}) = \alpha', \end{aligned} \tag{3.27}$$

where AD_i is computed under the $H_0 : A_i \sim U(0,1)$ and $A_1 \subset A_2 \cdots \subset A_m$. Apparently it is difficult to solve for c_i analytically. Instead, they can be fairly easily computed as the $(1-\alpha')\%$ quantiles of AD_i simulated according to Equation 3.27. For example, to derive c_2 according to the second line of Equation 3.27, we can simulate a random sample of A_1 which satisfies the condition $AD_1 < c_1$ and it is joined by another random sample $u_2 \sim U(0,1)$ forming A_2 for the computation of AD_2, and then c_2 is the $(1-\alpha')\%$ quantile of (pooled) multiple (e.g., 100,000) realisations of AD_2.

A preliminary analysis

Exposure Time Residuals (ETR) have not shown indicative sensitivity over model mis-specification in the analysis in previous sections, where the non-sequential approach is used. Here we consider simulated epidemics with a relatively significant number of cryptic exposures (i.e., unobserved exposures which are defined as exposures that have not yet become infectious in the observational period) such that the observations in the early stage of the epidemic are more likely to encapsulate more information of the transmission dynamics − cryptic exposures should correspond to "later" exposures and they play a lesser role in determining, for example, the latent period distribution.

We consider two epidemics with at least 10% cryptic exposures among all exposures. We consider a relatively restricted scenario where all exposures are known (in practice this may be achieved by performing diagnostic tests on subjects). Table 3.6 shows that the sequential approach based on ETR may be more sensitive to the mis-specification of the spatial kernel in such scenarios. Similar improvement of sensitivity is observed in the sequential approach based on ILR (Table 3.7).

Note that, as the critical values c_i are required to be re-computed when the sample size of the epidemic changes (i.e., to a classical observer, the testing result still depends on the random realisation of the epidemic size), this approach is not able to address this

Table 3.6: Values of $\pi(P(\widetilde{r}_1) < 0.05|y)$ and $\pi(T(\widetilde{r}_1) = 1|y)$, estimated from 1,500 posterior samples of ETR $\widetilde{r_1}$ for a simulated epidemic (about 10% cryptics out of all exposures) and analysed using two different model assumptions. Case A, the correct model structure; Case B, a mis-specified Cauchy-type spatial kernel.We consider a sequential testing at 6 levels of sample size 20, 50, 100, 200, 300 and 471 (471 is the number of exposures).

	Case A	*Case B*
$\pi(P(\widetilde{r}_1) < 0.05\|y)$	2%	9%
$\pi(T(\widetilde{r}_1) = 1\|y)$	3%	43%

Table 3.7: Values of $\pi(P(\widetilde{r}_2) < 0.05|y)$ and $\pi(T(\widetilde{r}_2) = 1|y)$, estimated from 1,500 posterior samples of ILR $\widetilde{r_2}$ for a simulated epidemic (about 15% cryptics out of all exposures) and analysed using two different model assumptions. Case A, the correct model structure; Case C, a mis-specified power-law spatial kernel. We consider a sequential testing at 6 levels of sample size 20, 50, 100, 200, 300 and 424 (424 is the number of exposures).

	Case A	*Case C*
$\pi(P(\widetilde{r}_2) < 0.05\|y)$	8%	14%
$\pi(T(\widetilde{r}_2) = 1\|y)$	8%	31%

rather subtle random-sample-size issue. These are also discussed in Chapter 5.

3.8 Discussion

It is well-known that the spatial transmission mechanisms are difficult to assess in practice yet have major implications for optimal control strategies. Studies of animals and plant diseases such as foot and mouth and citrus canker have cited the importance for selecting between a long-tailed spatial kernel versus a localised spatial kernel in devising most appropriate strategies of culling (Gottwald et al, 2002; Keeling et al, 2003; Ferguson et al, 2001). Therefore we believe that the methodology presented here, based on infection-link residuals (ILR), is a novel and potentially powerful tool for diagnosing mis-specification of a spatial kernel which can provide valuable insights to modellers in practice. Moreover, we remark that the principles introduced here should be readily extendable allowing the construction of analogous residuals for a wide range of processes included in models in ecology and epidemiology.

We also believe the approach offers several advantages over alternatives. Bayesian model assessment approaches, such as Bayes factors, are known to be sensitive to selection of prior distributions and are challenging computationally (Kass and Raftery, 1995; Han and Carlin, 2001). Moreover, they allow only relative comparison of com-

peting models, a disadvantage shared by information criteria measures such as the Deviance Information Criterion (DIC) (Spiegelhalter et al, 2002). The latter is also problematic when dealing with partially observed processes (Celeux et al, 2006), the norm in epidemiological studies, where the DIC is not uniquely defined. By contrast, the tests based on latent residuals offer an assessment of model discrepancy in absolute terms. Posterior predictive checks that utilise only partially observed data may be insensitive to the model choice (as shown in section 3.4 and section 3.6.2) even if summary statistics are appropriately chosen. A key feature of the proposed tests is that they can be easily embedded within any Bayesian analysis of a spatio-temporal system that makes use of data augmentation. Also, in contrast to other approaches such as posterior predictive checks, our method utilises the full posterior distribution of unobserved data and model parameters, and may offer a higher sensitivity to model mis-specifications. As it is common practice to conduct Bayesian analyses of partially observed epidemics using data augmentation supported by computational techniques such as Markov chain Monte Carlo methods (Ferguson et al, 2001; Catterall et al, 2012; Cook et al, 2007b; Gibson et al, 2006), the framework represents a potentially valuable addendum to the model-testing toolkit used in epidemiological and ecological studies.

Chapter 4

A new Bayesian computational method for the integrated analysis of epidemic and genetic data

4.1 Introduction

Individual-based, spatio-temporal stochastic models have proved to be extremely useful tools for analysis in epidemiological and ecological studies relating to transmission of diseases (Neri et al, 2014; Keeling et al, 2002). Models of this form can be formulated for situations where each site in the modelled population corresponds to an individual host such as a tree. For example, Citrus greening is a destructive citrus disease which has inflicted significant economic losses worldwide (Gottwald, 2010), and the understanding of its transmission dynamics in US state of Florida and the effectiveness of related control strategies is enhanced by a recent spatio-temporal epidemiological analysis (Parry et al, 2014). Other pathogens studied in this way include citrus canker (Neri et al, 2014). The models considered here can also be applied to represent a *metapopulation*, this being a set of well-defined spatially interacting subpopulations such as towns, farms, or spatial regions in space (Grenfell and Harwood, 1997). Metapopulation-level spatio-temporal epidemiological models have proved invaluable in studying several epidemic outbreaks including foot-and-mouth in livestock (Ster et al, 2009; Keeling et al, 2001) and measles in humans (Grenfell et al, 2001).

In spite of the considerable progress made in developing epidemiological models and associated methods of statistical inference, during a typical epidemic it may not be possible to observe the epidemiological data, such as contact structures and times

of infections, necessary to infer the detailed aspects of transmission including the transmission network. Genetic data on pathogens, which carry information on relatedness of different infection events, are increasingly becoming available and provide valuable insights during epidemic outbreaks. For example they can help to identify the transmission network, a key feature of interest, knowledge of which is relevant to quantifying superspreading events (Lloyd-Smith et al, 2005), to studying the evolutionary patterns of pathogens (Zhang et al, 1997; Leitner and Albert, 1999) and to designing and evaluating control measures (Ferguson et al, 2001).

Various statistical approaches have been proposed for the joint analysis of epidemiological and genetic data. Approaches that rely on reconstructing phylogenetic trees have been attempted in several scenarios (Shapiro et al, 2011; Grenfell et al, 2004; Rambaut et al, 2008). However, as noted in Jombart et al (2010) these approaches may be inappropriate when the sampled sequences contain both ancestors and their descendants, which is particularly the case during the early stages of an epidemic (see also discussion concerning these approaches in Chapter 1). On the other hand, substantial progress has been made by considering combining genetic data with explicitly constructed transmission trees (Ypma et al, 2012; Morelli et al, 2012; Ypma et al, 2013; Jombart et al, 2014; Cottam et al, 2008). These current state-of-the-art approaches make use of approximations to analyse epidemiological and evolutionary processes jointly either by considering pseudo-likelihoods (Morelli et al, 2012; Jombart et al, 2014) that only take into account observed sequences, or by assuming sequence combinations that exhibit the minimum amount of mutation necessary to explain the subtrees of transmission (Ypma et al, 2012). Such approximations, which both implicitly assume independence between any two subsets of the transmission network, greatly reduce the inherent computational challenges in joint analysis of epidemiological and evolutionary process and facilitate statistical inference when the transmission network is of primary interest. Other model parameters, such as the transmission rate and the dispersal kernel, are also of significant importance for predicting dynamics and management of epidemics (Parry et al, 2014; Ster et al, 2009; Ferguson et al, 2001), however, have not received the same attention as the transmission network, and their estimation may not be robust to such approximations. Further research on the integration of epidemiological and genetic data in the context inference of both the transmission network and the transmission dynamics is certainly warranted.

As noted by Morelli et al (2012), a more exact account of the joint process is hindered by an inherent and key unresolved challenge, that is, to impute effectively the *unobserved* and transmitted pathogen sequences, which typically requires a very high-dimensional model space. In this chapter we extend previous research by formulating a framework in which unobserved, transmitted pathogen sequences are explicitly

included. Consequently, the joint process is more accurately captured and model parameters governing the transmission dynamics, including the transmission network, may be estimated in a more valid manner. Specifically, we aim to impute effectively the unobserved transmitted sequences and to relax some of the restrictive assumptions made currently in the literature. We will allow greater uncertainty in times of infection, which may not be generally available, and will consider epidemics with unknown numbers of 'clusters' where a cluster is set of infections arising from a single primary infection. Note that, to include multiple-cluster scenarios, a transmission network should be technically called a transmission *graph* instead of the term transmission *tree* typically used in the literature (Ypma et al, 2012; Morelli et al, 2012). We note that not only does the imputation of the unobserved sequences enable a more accurate account of transmission dynamics, it also offers the key advantage that unsampled exposures (i.e., infected hosts without observed sequence samples), whose role in transmission dynamics should unequivocally be taken into account, can be naturally accommodated, in contrast to existing approaches in the literature.

We show that unobserved transmitted sequences can be imputed effectively and the transmission dynamics can be reasonably recovered in the presence of exposures without observed sequences. Using the proposed framework, we study comprehensively the role of genetic data in understanding the transmission dynamics. Specifically, we show that increased availability of genetic data can aid the estimation of epidemiological model parameters, and we demonstrate the value of partial genomic data in quantifying outbreaks and have important implications for sampling designs of future studies. Moreover, we demonstrate that genetic data may enhance our ability to detect mis-specification of the spatial transmission mechanism when they are used in combination with the residual methods of Chapter 3. The proposed framework is subsequently applied to analyse a localised spread of foot-and-mouth disease virus in the UK and we show that understanding of the transmission dynamics can be greatly enhanced. Some of the results in this chapter are under peer-review for publication (Lau et al, 2014a).

Specifically, in this chapter we aim to achieve the following:

- to devise a statistically sound Bayesian framework which facilitates the integration of epidemiological and genetic data; specifically, we demonstrate how the unobserved data, including the transmitted sequences, can be effectively imputed so that the transmission dynamics of the joint epidemic and evolution process can be accurately recovered and also any unsampled infected hosts can be naturally accommodated in the analysis;
- to consider the general scenario of a multiple-cluster transmission graph (i.e.,

presence of multiple primary infections) and demonstrate that it can be adapted to the single-cluster transmission graph;

- to characterise the importance of genetic data in the inference of transmission dynamics, specifically, in the estimation of the transmission graph, epidemiological parameters and the assignment of exposures to the clusters; and to investigate how genetic data may facilitate model assessment using methods developed in Chapter 3.
- to demonstrate the effectiveness of the framework using simulated data and to apply it to a dataset describing the spread of foot-and-mouth disease virus in a UK region in 2001.

4.2 Model and methods

4.2.1 The stochastic epidemic process

Similar to Chapter 3, we consider a broad class of spatio-temporal stochastic models exemplified by the SEIR epidemic model. Here we recall the notation to facilitate the reading of this chapter. Suppose that we have a spatially distributed population indexed by 1, 2, ... Denote by $\xi_S(t)$, $\xi_E(t)$, $\xi_I(t)$ and $\xi_R(t)$ the set of indices of individuals who are in class S, E, I and class R respectively at time t and let $S(t)$, $E(t)$, $I(t)$ and $R(t)$ be the respective numbers in these classes at time t.

The *transmission graph* is also explicitly modelled in the epidemic process. To be specific, an individual $j \in \xi_S(t)$ becomes exposed via primary infection with stochastic rate α and from an infection $i \in \xi_I(t)$ with rate $\beta K(d_{ij}; \kappa)$. Sources of infection are assumed to act independently of each other and combine so that the overall probability of j becoming infected during $[t, t + dt)$ is given by

$$r(j, t, dt) = (\alpha + \beta \sum_{i \in \xi_I(t)} K(d_{ij}; \kappa))dt + o(dt). \tag{4.1}$$

We use a $Gamma(a, b)$ parameterised by the shape a and scale b for the random time x spent in class E. For the random time x spent in class I we use a $Weibull(\gamma, \eta)$ parameterised by the shape γ and scale η. All sojourn times are assumed independent of each other given the model parameters.

We define a *cluster* to be a set of infections arising from a single primary infection. Note that the magnitude of primary infection rate α is the determinant factor for

the number of primary cases and hence the number of clusters of the transmission graph.

4.2.2 The stochastic evolutionary process

The evolutionary process of the pathogen is modelled at the level of nucleotide substitutions (see details in Chapter 2). It is assumed that the nucleotide substitution process is conditionally independent of the epidemic process. We assume that there is only a single dominating strain at each exposed individual at any time point. Upon exposure, the newly exposed individual is assumed to be infected by a single dominant strain from the source individual which is subject to a continuous-time evolutionary process described below. Nucleotide bases at different positions of a sequence are assumed to evolve independently.

Taking RNA viruses as an example, we let $\omega_N = \{A, C, G, U\}$ be the set of nucleotide bases. We consider the Kimura model (see also Chapter 2) in which a nucleotide base $x \in \omega_N$ mutates to a nucleotide base $y \in \omega_N$ within a time duration $\triangle t$ with probability

$$p_{\mu_1,\mu_2}(y|x, \triangle t) = 0.25 + 0.25e^{-4\mu_2\triangle t} + 0.5e^{-2(\mu_1+\mu_2)\triangle t}, \quad \text{for } x = y, \tag{4.2a}$$

$$p_{\mu_1,\mu_2}(y|x, \triangle t) \tag{4.2b}$$

$$= \begin{cases} 0.25 + 0.25e^{-4\mu_2\triangle t} - 0.5e^{-2(\mu_1+\mu_2)\triangle t}, & \text{for } x \neq y \text{ and it is a transition,} \\ 0.25 - 0.25e^{-4\mu_2\triangle t}, & \text{for } x \neq y \text{ and it is a transversion,} \end{cases} \tag{4.2c}$$

where μ_1 and μ_2 are the rates of transition and transversion respectively.

4.2.3 Modelling background pathogen and multiple clusters

Current state-of-the-art approaches to joint inference of epidemiological and genetic data mostly assume that primary cases (other than the index case) have no role in the transmission network (Morelli et al, 2012; Ypma et al, 2012) and hence focus on epidemics characterised by a single cluster, or rely on *ad hoc* approaches which identify probable genetic outliers as the primary cases without explicitly modelling the background infection process (Jombart et al, 2014). In the following sections, we first consider the general scenario where primary cases are modelled explicitly so that

multiple clusters in the transmission graph can be accommodated. Subsequently we demonstrate that the single-cluster scenario can be readily accommodated.

To consider the general multiple-cluster scenario, it will be necessary to describe the distribution of any background sequences of the pathogen so that a background infection and a secondary infection become distinguishable using the Bayesian computational procedures used here. In order to represent both single and multiple clusters scenarios, we consider a given *master sequence*, G_M, which may be considered as a universal antecedent of any background sequence, with each nucleotide base of a background sequence having a probability p of differing from the same-position base in the master sequence (if selected for a change a different base is chosen with equal probability from the three alternatives). Deviations are assumed to be independent across sites. We note that the concept of having a master sequence representing the background population is similar to the concept of deriving a theoretical *consensus sequence* (Turner, 2005) from the alignment of multiple sequences.

4.2.4 Likelihood

Consider a population of size N and pathogen sequences comprising n bases. The epidemic is observed between time $t = 0$ and $t = t_{max}$. Let χ_S denote the set of individuals remaining in class S by t_{max}, and denote χ_E, χ_I and χ_R the set of individuals who have ever gone through class E, class I and class R by t_{max} respectively. Also, let $\mathbf{E} = (\ldots, E_j, \ldots)$ denote the exposure times for $j \in \chi_E$, $\mathbf{I} = (\ldots, I_j, \ldots)$ denote the times of becoming infectious for $j \in \chi_I$ and $\mathbf{R} = (\ldots, R_j, \ldots)$ denote the times of recovery for $j \in \chi_R$. The cumulative distribution functions corresponding to the sojourn times in class E and class I are denoted as F_E and F_I respectively.

Furthermore, let $G_{\cdot j} = (G_{1,j}, G_{2,j,}, \ldots, G_{m_j,j})$ denote the m_j sequences on individual $j \in \chi_E$ with corresponding *sequencing times* $t_{\cdot j} = (t_{1,j}, t_{2,j,}, \ldots, t_{m_j,j})$ sorted in ascending order. Note that these sequencing times are the times where model formulation needs to account for, and as such they include the observed *sampling time* t_j^s and unobserved time of exposure E_j and unobserved times of j infecting any other individuals. And $\boldsymbol{G} = (G_{\cdot 1}, \ldots, G_{\cdot j}, \ldots)$ specifies the complete nucleotide data where only the sampled sequences are observed. Moreover, transmission graph $\boldsymbol{\psi}$ specifies the source of infection ψ_j for any individual $j \in \chi_E$. The event of individual i infecting individuals j and k and the sampling of sequences on these individuals is shown in Figure 4.1 to help illustrate the notations above. It is also worth noting that among these events only the sampling times t_i^s, t_j^s, t_k^s and the corresponding sequence samples (coloured grey) are observed. Note that for any exposed individual j, the first sequencing time $t_{1,j}$ is equal to the exposure time E_j and the first sequence $G_{1,j}$ is

identical to the sequence present on the source ψ_j at E_j.

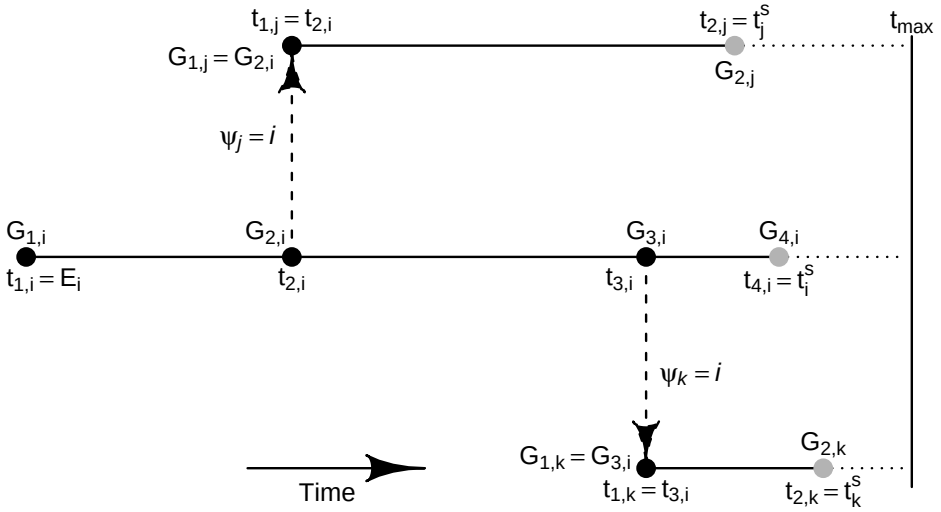


Figure 4.1: The event of individual i infecting individuals j and k (dashed arrows) and the sampling of sequences on these individuals. Among these events only the sampling times t_i^s, t_j^s, t_k^s and the corresponding sequence samples (coloured grey) are observed and other unobserved quantities are to be estimated (see later). Solid circles represent the sequences at respective time points. Possible events on dotted lines are not shown. Note that in our inference we do not demand that all exposures have an observed sequence sample.

Likelihood with multiple clusters

In the general multiple-cluster scenario, with the complete data $\boldsymbol{z} = (\mathbf{E}, \mathbf{I}, \mathbf{R}, \mathbf{G}, \boldsymbol{\psi})$ and model parameters $\boldsymbol{\theta} = (\alpha, \beta, a, b, \gamma, \eta, \kappa, \mu_1, \mu_2, p)$, we can express the likelihood as

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}) = & \prod_{j \in \chi_E^{-1}} p(\psi_j, j | E_j) \times e^{-q_j} \times \prod_{j \in \chi_S} e^{-q_{T_j}} \\ & \times \prod_{j \in \chi_I} f_E(I_j - E_j; a, b) \times \prod_{j \in \chi_R} f_I(R_j - I_j; \gamma, \eta) \\ & \times \prod_{j \in \chi_{E \setminus I}} \{1 - F_E(t_{max} - E_j; a, b)\} \times \prod_{j \in \chi_{I \setminus R}} \{1 - F_I(t_{max} - I_j; \gamma, \eta)\} \\ & \times \prod_{j \in \chi_E} g(G_{2,j}, \ldots, G_{m_j,j} | t_{\cdot j}, \psi_j, G_{1,j}) \times \prod_{j \in \chi_E} h(G_{1,j} | \psi_j) \end{aligned} \tag{4.3}$$

with the contribution to the likelihood arising from the infection of j by ψ_j being

$$p(\psi_j, j|E_j) = \begin{cases} \alpha, & \text{if individual } j \text{ is a primary case,} \\ \beta K(d_{\psi_j j}; \kappa), & \text{if } \psi_j \in \chi_{I\setminus R} \text{ at time } E_j. \end{cases} \tag{4.4}$$

Also,

$$q_j = \int_{t=0}^{E_j} \{\alpha + \sum_{i \in \xi_I(t)} \beta K(d_{ij}; \kappa)\} dt, \tag{4.5}$$

and

$$q_{T_j} = \int_{t=0}^{t_{max}} \{\alpha + \sum_{i \in \xi_I(t)} \beta K(d_{ij}; \kappa)\} dt, \tag{4.6}$$

with χ_E^{-1} being χ_E excluding the index case.

A more accurate account of the evolutionary process We have

$$g(G_{2,j}, \ldots, G_{m_j,j}|t_{\cdot j}, \psi_j, G_{1,j}) = \prod_{i=1}^{n} \prod_{k=1}^{m_j - 1} p_{\mu_1,\mu_2}(G^i_{k+1,j}|G^i_{k,j}, \triangle t = t_{k+1,j} - t_{k,j}) \tag{4.7}$$

giving the conditional probability of the sequences with $p_{\mu_1,\mu_2}(\cdot)$ defined in Equation 4.2 (where $G^i_{k,j}$ denotes the nucleotide base on i^{th} position of k^{th} sequence on individual j). It is worth noting that $g(G_{2,j}, \ldots, G_{m_j,j}|t_{\cdot j}, \psi_j, G_{1,j})$ accounts for unobserved sequences and this allows exact calculation of the likelihood associated with the assumed evolutionary process (also see Figure 4.1). Morelli et al (2012) consider an *ad hoc* pseudo-likelihood approach to approximating $\prod_{j \in \chi_E} g(G_{2,j}, \ldots, G_{m_j,j}|t_{\cdot j}, \psi_j, G_{1,j})$ for the evolutionary process, without accounting for the unobserved sequences and (hence) the dependence between any given subtrees. For example, for the sequence of events in Figure 4.1, instead of taking into account the unobserved transmitted sequences, they consider pairs of observed sequence samples and the corresponding *earliest divergent times* (see below) where

$$\begin{aligned} \prod_{z \in \{i,j,k\}} g(G_{2,z}, \ldots, G_{m_z,z}|t_{\cdot z}, \psi_z, G_{1,z}) = \prod_{l=1}^{n} & p_{\mu_1,\mu_2}(G^l_{4,i}|G^l_{2,j}, \triangle t = |t^s_j - E_j| + |t^s_i - E_j|) \\ & \times p_{\mu_1,\mu_2}(G^l_{4,i}|G^l_{2,k}, \triangle t = |t^s_k - E_k| + |t^s_i - E_k|) \\ & \times p_{\mu_1,\mu_2}(G^l_{2,j}|G^l_{2,k}, \triangle t = |t^s_k - E_j| + |t^s_i - E_j|) \end{aligned} \tag{4.8}$$

and the respective earliest divergent times for the three pairs are E_j, E_k and E_j. Without accounting for the unobserved sequences, this approach does not account for the dependence among the three terms in the equation above.

In contrast, our approach explicitly takes into account the subtrees dependence, by accounting for the unobserved sequences, and is therefore able to describe adequately the assumed evolutionary process.

Background pathogens The expression $h(G_{1,j}|\psi_j)$ essentially depicts the distribution of a sequence for a primary case. Here we model the probability of a background sequence as

$$h(G_{1,j}|\psi_j) = \begin{cases} (\frac{p}{3})^{l_j}(1-p)^{n-l_j}, & \text{if individual } j \text{ is a primary case,} \\ 1, & \text{if } \psi_j \in \chi_I, \end{cases} \tag{4.9}$$

where p is the probability that a base of $G_{1,j}$ is different from the base at the corresponding position at the given *master sequence* G_M and l_j is the total number of different bases. The term $\frac{1}{3}$ reflects the assumption that a base can be randomly chosen from the set $\omega_N \setminus G_M^i$, where G_M^i is the nucleotide base on i^{th} position of the master sequence.

Likelihood with single cluster

The framework in Section 4.2.4 can be easily adapted to a single-cluster scenario (Ypma et al, 2012; Morelli et al, 2012) by disregarding the process of generating the background sequences (i.e., it is not necessary to consider the master sequence G_M and parameter p in this scenario). Specifically, we have

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}) = & \prod_{j \in \chi_E^{-1}} p(\psi_j, j|E_j) \times e^{-q'_j} \times \prod_{j \in \chi_S} e^{-q_{T_j}} \\ & \times \prod_{j \in \chi_I} f_E(I_j - E_j; a, b) \times \prod_{j \in \chi_R} f_I(R_j - I_j; \gamma, \eta) \\ & \times \prod_{j \in \chi_{E \setminus I}} \{1 - F_E(t_{max} - E_j; a, b)\} \times \prod_{j \in \chi_{I \setminus R}} \{1 - F_I(t_{max} - I_j; \gamma, \eta)\} \\ & \times \prod_{j \in \chi_E} g(G_{2,j}, \ldots, G_{m_j,j}|t_{\cdot j}, \psi_j, G_{1,j}) \end{aligned} \tag{4.10}$$

4.2.5 Bayesian inference and MCMC

The success of an MCMC algorithm is often determined by the choice of the *proposal distributions* for model parameters in Metropolis-Hastings algorithms and also the correct specifications of the respective *acceptance probabilities*, so that the distribution of the state of the Markov chain *converges* to the target posterior distribution with sufficient speed for samples from the chain to be used to explore the target distribution (Chib and Greenberg, 1995) (see detailed discussion in Chapter 2).

In our application, $\boldsymbol{z}$ involves both partially observed epidemic and sequence data and the inference is accordingly more challenging than that of analysing epidemic data only, as it involves exploring the very high-dimensional model space constituted from the large numbers of combinations of transmission graphs and sequences (Morelli et al, 2012). Standard MCMC algorithms, such as the single-step Metropolis-Hastings make updates to a single model quantity at any time. However, for high-dimension problems similar to the problem in this work, well-designed joint proposal schemes for the model quantities are generally challenging, but necessary for obtaining a well-converged Markov chain that can explore efficiently the joint posterior distribution of the model quantities. For instance, in the sampling of an unobserved exposure time, a naive algorithm may update the time leaving the corresponding transmitted sequence unchanged. This may lead to a very low acceptance probability for the proposed change, such that the domain of the exposure time is not explored efficiently.

A crucial research gap, therefore, for the joint inference of epidemic and molecular evolution processes is to devise a statistically sound, and computationally efficient algorithm for the imputation of unobserved sequences and the transmission graph $\boldsymbol{\psi}$. In this section we describe how the unobserved $\boldsymbol{\psi}$ and the unobserved sequences in $\boldsymbol{G}$ may be sampled together with the unobserved exposure times $\boldsymbol{E}$, which is being the key challenge in devising a suitable algorithm. Other more standard elements of the MCMC algorithm are referred to Section 4.2.6 (see later). Computer experiments and mathematical arguments to validate the methods are also presented in Section 4.7.

Table 4.1: Notation used in Section 4.2.5.

Notation	Description
E_j	The exposure time of site j.
$t_{\cdot j} = (t_{1,j}, \ldots, t_{m_j,j})$	The vector containing m_j relevant *sequencing times* on exposed site $j \in \chi_E$.
$G_{\cdot j} = (G_{1,j}, \ldots, G_{m_j,j})$	The vector containing corresponding *sequences* at times in the vector $t_{\cdot j}$.
t_j^s	The *observed sampling time* in the vector $t_{\cdot j}$.
$G_{1,j}^k$	The *nucleotide base* at k^{th} position in the sequence $G_{1,j}$.
G_M and G_M^k	The background master sequence and its k^{th}-position nucleotide base.
ψ_j	The source of infection for exposed site j.

Joint Sampling of the Exposure Time E_j and the Corresponding Sequence $G_{1,j}$

Assuming for now that the source of infection ψ_j is unchanged, and given the current exposure time E_j for individual j and the corresponding sequence $G_{1,j}$, we first propose a new exposure time $E_j^{'}$ using a standard approach (see details in Section 4.2.6). Here we describe in detail how a suitable candidate for the corresponding sequence $G'_{1,j}$ can be simultaneously proposed. To facilitate reading of the current and following sections some key notation is summarised in Table 4.1.

The key idea is to propose a new sequence at $E_j^{'}$ which has plausible proximity to a *nearest past sequence* G_p and a *nearest future sequence* G_f relative to $E_j^{'}$. Throughout *past* and *future* are defined self consistently in terms of the direction of $\Delta E_j = E_j^{'} - E_j$. Therefore, if $E_j^{'}$ is before E_j, G_p will be at a later (absolute) time than G_f. We choose G_p and G_f by taking account of the sequences both from individual j and ψ_j. Denoting t_p and t_f as the sequencing times for G_p and G_f respectively, we have notationally

$$t_p = \arg\min\{\left|t - E_j^{'}\right| : t \in t_{\cdot j} \cup t_{\cdot \psi_j} | \operatorname{sgn}(t - E_j^{'}) \neq \operatorname{sgn}(\Delta E_j)\} \tag{4.11}$$

and

$$t_f = \arg\min\{\left|t - E_j^{'}\right| : t \in t_{\cdot j} \cup t_{\cdot \psi_j} and \operatorname{sgn}(t - E_j^{'}) = \operatorname{sgn}(\Delta E_j)\}, \tag{4.12}$$

where *sgn* is the *signum function* defined as follows:

$$sgn(t) = \begin{cases} -1, & \text{if } t < 0, \\ 0, & \text{if } t = 0, \\ +1, & \text{if } t > 0. \end{cases} \tag{4.13}$$

G_p (or G_f) is taken to be the corresponding sequence on individual j whenever t_p (or t_f) is in both $t_{\cdot j}$ and $t_{\cdot \psi_j}$. This is illustrated in Figure 4.2 where a new exposure time $E_j^{'}$ for individual j from Figure 4.1 is proposed. Here t_p and t_f are taken to be the current exposure time E_j and $t_{3,i}$ respectively. Then, by definitions, the corresponding sequences at t_p and t_f are $G_p = G_{1,j}$ and $G_f = G_{3,i}$ respectively. It is noted that $G_{2,i}$ and $t_{2,i}$ are also simultaneously updated.

Given the nucleotide base G_p^i and G_f^i at the i^{th} position and time apart $\Delta t_p = |t_f - t_p|$, by conditioning on at most one change occurring in the period Δt_p, and assuming a linear relationship between the probability of change and the time duration, we

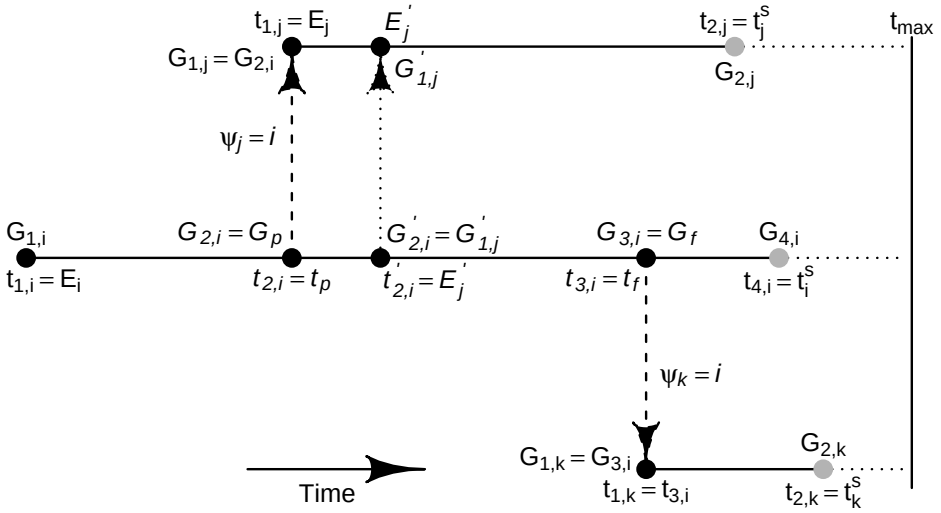


Figure 4.2: Illustration of the selection t_p (and the corresponding past sequence G_p) and t_f (and the corresponding past sequence G_f) (also see main text).

have

$$G'^{i}_{1,j} = \begin{cases} G^i_f, & \text{with probability } p_f = \frac{\left|E'_j - t_p\right|}{\Delta t_p}, \\ G^i_p, & \text{with probability } 1 - p_f. \end{cases} \tag{4.14}$$

As $\triangle t_p$ is generally small, we allow only one change during the time interval for a particular nucleotide base, which is not entirely consistent with the assumption of a Markov process, in order to thoroughly explore the domain of $\boldsymbol{G}$, $\boldsymbol{G}$ is also updated independently of the exposure times (see Section 4.2.6).

It is noted that although t_p and G_p are always well-defined, as the corresponding set in Equation 4.11 is always non-empty and contains E_j. t_f and G_f may be undefined as the corresponding set in Equation 4.12 can be empty. If G_f is not well-defined, we consider a proposal of $G'_{1,j}$ according to the mechanism defined in Equation 4.2 such that G^i_p mutates to $G'^{i}_{1,j} = y$ with probability

$$P(G'^{i}_{1,j} = y | G^i_p, \triangle t = \left|E'_j - t_p\right|) = p_{\mu_1,\mu_2}(G'^{i}_{1,j} = y | G^i_p, \triangle t = \left|E'_j - t_p\right|). \tag{4.15}$$

In the case of $\psi_j \notin \chi_I$ and when $G_{2,j}$ is not available, and thus the newly proposed sequence is not constrained by any secondary sequence, any sensible proposal distribution must take account of assumptions regarding the background sequence. Specifically, $G'^{i}_{1,j}$ has a probability $1-p$ of matching the corresponding site G^i_M in the master sequence G_M and a probability p of being different (in which case a different base is randomly drawn from the set $\omega_N \setminus G^i_M$).

Lastly, the updating of the current data $\boldsymbol{z}$ to $\boldsymbol{z}'$ is accepted with a M-H acceptance probability (see later). By sequentially applying this algorithm to all exposures $j \in \chi_E$, $\boldsymbol{E}$ and $\boldsymbol{G}$ can be jointly updated.

Joint sampling of ψ_j and $G_{1,j}$ and E_j

Denote t_u as the upper limit of $E_j^{'}$ (see Section 4.2.6) and $\omega_\psi = \{i \in \chi_I | I_i \leq t_u, i \neq \psi_j\}$ as the set of candidates for a new source of infection $\psi_j^{'}$. Utilising spatial connectivity, an infective $i \in \omega_\psi$ is chosen to be $\psi_j^{'}$ with probability

$$s_{ij} \propto K(d_{ij}; \kappa). \tag{4.16}$$

Note that for the multiple-cluster scenario the primary infection can be accommodated by adding a permanent infectious challenge α to individual j. After the sampling of $\psi_j^{'}$, $E_j^{'}$ can be subsequently sampled (see Section 4.2.6) and hence $t_{\cdot j}^{'}$ and $t_{\cdot \psi_j^{'}}^{'}$ are also updated.

The corresponding newly proposed sequence $G_{1,j}^{'}$ is different from Section 4.2.5 as E_j and $G_{1,j}$ become irrelevant when the source of infection also changes. In the case of a new source $\psi_j^{'} \in \chi_I$ we define

$$t_p = \arg\min\{\left|t - E_j^{'}\right| : t \in t_{\cdot \psi_j^{'}}^{'} and t < E_j^{'}\}, \tag{4.17}$$

where $t_{\cdot \psi_j^{'}}^{'}$ are the updated sequencing times on $\psi_j^{'}$ (which is simultaneously updated after the updates of E_j and ψ_j) and then we can identify the respective sequence G_p. Also, we define

$$t_f = \arg\min\{\left|t - E_j^{'}\right| : t \in t_{\cdot j}^{'} \cup t_{\cdot \psi_j^{'}}^{'} and t > E_j^{'}\}, \tag{4.18}$$

where $t_{\cdot j}^{'}$ are the updated sequencing times on j. Note that $t_f > t_p$ always holds in the definitions in this case. G_f is taken to be the corresponding sequence on individual j whenever t_f is in both $t_{\cdot j}^{'}$ and $t_{\cdot \psi_j^{'}}^{'}$. $G_{1,j}^{'}$ is then sampled according to Equation 4.14. Similarly, $G_{1,j}^{'}$ is sampled according to Equation 4.15 when G_f is not well-defined.

In the case of $\psi_j^{'} \notin \chi_I$, we let $G_p = G_{2,j}$ and sample $G_{1,j}^{'}$ according to Equation 4.15; if $G_{2,j}$ is not available, $G_{1,j}^{'i}$ is sampled according to the assumption of the background sequences as shown in Section 4.2.5. Similarly, the updating of the current data $\boldsymbol{z}$ to $\boldsymbol{z}'$ is accepted with a M-H acceptance probability (see later).

Sampling of cryptic exposures

As methods described in the current literature do not impute the unobserved transmitted sequences, unobserved exposures cannot easily be accommodated by them. Denote ω_C as the set of exposed individuals who do not have an observed sample and are not in χ_I. We refer to $j \in \omega_C$ as to a *cryptic* exposure. We incorporate j in our framework by imputing the sequence transmitted to j. Allowing cryptic exposures requires a 'swap' of individuals between the sets ω_C and χ_S and a transmitted sequence needs to be imputed when an individual from χ_S moves to ω_C. After the individual to be swapped has been proposed, the sequence is imputed in a same manner as Section 4.2.5. The acceptance probability is similar to that in Section 4.2.5 with an additional term which accounts for the 'swapping' probability (Gibson and Renshaw, 1998). See also details in Chapter 2 for a general discussion for RJMCMC and a specific application in 3.3.8.

Initialisation of $\boldsymbol{\psi}$

When only a subset of individuals $j \in \chi_E$ have an observed sequence sample, the choice of the starting value of $\boldsymbol{\psi}$ becomes important for the rate of convergence of the Markov chain. In this case, we sample the starting value $\boldsymbol{\psi}_0$ from the marginal posterior distribution of $\boldsymbol{\psi}$, $P_e(\boldsymbol{\psi})$, obtained from only fitting the epidemic model to the epidemic data by adopting a standard Metropolis-Hastings algorithm. Effectively, we set $g(\cdot) = h(\cdot) = 1$ in Equation 4.3 and we need not impute the unobserved sequences. As we can see in Section 4.3, samples from $P_e(\boldsymbol{\psi})$ provide a good approximation to the true $\boldsymbol{\psi}$ and hence can be used as a reasonable starting point in the joint analysis of epidemic and genetic data.

Acceptance probabilities

The acceptance probability of a proposed parameter value θ_i' with current value θ_i is

$$p_a = \min\{1, \frac{L(\boldsymbol{\theta}^{'}; \boldsymbol{z})}{L(\boldsymbol{\theta}; \boldsymbol{z})} \times \frac{P(\theta_i')}{P(\theta_i)} \times \frac{q(\theta_i|\theta_i')}{q(\theta_i'|\theta_i)}\} \tag{4.19}$$

where $P(\theta_i)$ is the *prior distribution* of θ_i and $q(\theta_i'|\theta_i)$ *proposal distribution* of θ_i' given the current value θ_i. Acceptance probability in proposing a component of augmented data $\boldsymbol{z}_i'$ is similar by setting $\boldsymbol{\theta}$ constant. In most of the cases, q is symmetric in the reverse direction and hence the *proposal ratio* (e.g., $\frac{q(\theta_i|\theta_i')}{q(\theta_i'|\theta_i)}$) reduces to 1, which simplifies the problem. However, when the proposal density is less straightforward, one has to work out the proposal ratio.

We describe in details of computing the proposal ratio for the joint sampling of $E_j^{'}$ and $G_{1,j}^{'}$ for Section 4.2.5. As an illustration, we consider only the case where G_p and G_f are both defined. We have to compute the *forward proposal probability* (i.e., the denominator)

$$q(E_j^{'}, G_{1,j}^{'}|E_j, G_{1,j}) = q_1(E_j^{'}|\boldsymbol{E}) \times q_2(G_{1,j}^{'}|E_j^{'}, \boldsymbol{E}, \boldsymbol{G}) \tag{4.20}$$

and the *backward proposal probability* (i.e., the numerator)

$$q(E_j, G_{1,j}|E_j^{'}, G_{1,j}^{'}) = q_1(E_j|\boldsymbol{E}') \times q_2(G_{1,j}|E_j, \boldsymbol{E}', \boldsymbol{G}'). \tag{4.21}$$

ψ_j is unchanged so that the domain of E_j and $E_j^{'}$ is the same and we have

$$\frac{q_1(E_j|\boldsymbol{E}')}{q_1(E_j^{'}|\boldsymbol{E})} = 1. \tag{4.22}$$

We also have

$$q_2(G_{1,j}^{'}|E_j^{'}, \boldsymbol{E}, \boldsymbol{G}) = p_f^{m_f} \times (1 - p_f)^{n-m_f} \tag{4.23}$$

where m_f is the number of nucleotide on G_f which are the same as the corresponding one (i.e., same position) on $G_{1,j}^{'}$. $q_2(G_{1,j}|E_j, \boldsymbol{E}', \boldsymbol{G}')$ is similarly computed by considering the reverse direction - in particular, we have to re-define G_p and G_f as the direction of change of time is reversed. The proposal ratio for Section 4.2.5 can be easily computed in a similar manner. However, we have to give consideration to the difference in the domains between E_j and $E_j^{'}$ and the ratio of probabilities of proposing ψ_j and ψ_j' according to details given in Section 4.2.5.

4.2.6 Other details of the MCMC algorithm

In this section we give other details of the algorithm which are more standard and not described in last section.

Sampling of E_j

For Section 4.2.5, $E_j^{'}$ is proposed as a random draw

$$E_j^{'} \sim U(t_l, t_u). \tag{4.24}$$

If j is a primary infection we have

$$t_l = 0 \tag{4.25}$$

and

$$t_u = \begin{cases} \min\{t_j^s, I_j\}, & \text{if } j \text{ has an observed sequence sample (at } t_j^s) \text{ and } j \in \chi_I, \\ t_j^s, & \text{if } j \text{ has an observed sequence sample and } j \notin \chi_I, \\ I_j, & \text{if } j \text{ has no observed sequence sample and } j \in \chi_I, \\ t_{max}, & \text{if } j \text{ has no observed sequence sample and } j \notin \chi_I. \end{cases} \tag{4.26}$$

In the case of $\psi_j \in \chi_I$, we have

$$t_l = I_{\psi_j} \tag{4.27}$$

and

$$t_u = \begin{cases} \min\{t_j^s, I_j, R_{\psi_j}\}, & \text{if } j \text{ has an observed sequence sample (at } t_j^s) \text{ and } j \in \chi_I, \\ \min\{t_j^s, R_{\psi_j}\}, & \text{if } j \text{ has an observed sequence sample and } j \notin \chi_I, \\ \min\{I_j, R_{\psi_j}\}, & \text{if } j \text{ has no an observed sequence sample and } j \in \chi_I, \\ R_{\psi_j}, & \text{if } j \text{ has no an observed sequence sample and } j \notin \chi_I. \end{cases} \tag{4.28}$$

When $\psi_j \notin \chi_R$, Equation 4.28 reduces to Equation 4.26. For Section 4.2.5, $E_j^{'}$ is proposed in the same manner with ψ_j now being replaced by $\psi_j^{'}$. The reader is reminded that the sampling of $E_j^{'}$ is only part of the joint sampling procedures in Section 4.2.5.

Sampling of I_j

To incorporate uncertainty in infectious times, I_j is assumed to be known within a range - let t_o denote a known time (e.g., the symptoms onset time in practice) such that I_j is randomly drawn within a range $t_o \pm D$. For simulation studies, we assume t_o to be the true I_j and $D = 0.6$. Considering also E_j and R_j, we have $I_j^{'}$ sampled as a random draw between t_a and t_b where

$$t_a = \max\{E_j, t_o - D\} \tag{4.29}$$

and

$$t_b = \min\{R_j, t_o + D\}. \tag{4.30}$$

R_j is replaced by t_{max} if $j \notin \chi_R$. Note that from the sampling perspective, we are effectively assuming that an observation (i.e., a diagnostic test for infectiousness) is taken at $t_o - D$ and at $t_o + D$, the first being negative and the second positive.

Sampling of $G_{1,j}$

Other than the joint sampling of $E_j^{'}$ and $G_{1,j}^{'}$ which enables us to explore the joint model space more effectively, separate updating of $G_{1,j}^{'}$ is necessary for a thorough exploration of the domain of $\boldsymbol{G}$. We implement a simple updating algorithm for sampling $G_{1,j}^{'}$ - for an individual $j \in \chi_E$, each nucleotide base $G_{1,j}^{'i}$ is sampled uniformly from the set $\omega_N = \{A, C, G, T\}$.

Sampling of $\boldsymbol{\theta}$

Each parameter in $\boldsymbol{\theta} = (\alpha, \beta, a, b, \gamma, \eta, \kappa, p, \mu_1, \mu_2)$ is updated sequentially with a standard random-walk Metropolis-Hastings algorithm (see details in Chapter 2 and in particular 3.3.8).

Sampling of G_M and the initial value

The master sequence G_M determines the source sequence for a particular cluster and the choice of its initial value in MCMC is crucial. Specifically, we choose the first observed sample in the population as the initial value. We implement a simple updating algorithm similar to the updating of $G_{1,j}$ - each nucleotide base $G_M^{'i}$ is sampled uniformly from the set $\omega_N \setminus G_M^i$.

Initial values of other parameters

In the application to FMDV (Section 4.6), we use $\alpha = 0.0002$, $\beta = 3.0$, $a = 2.0$, $b = 2.0$, $\mu = 8.0$, $\kappa = 0.1$, $\mu_1 = 0.0001$ and $\mu_2 = 0.00005$ as the initial values. For simulation studies (Section 4.3) each parameter in $\boldsymbol{\theta}$ is initialised to be one half of its true value. An initial value of I_j, $j \in \chi_I$, is randomly drawn within a range $t_o \pm D$. Set the initial χ_E be χ_I (i.e., no cryptic exposures). Let $\chi_\psi = \{j \in \chi_I | I_j \leq t_u\}$ be the set of possible sources for a particular individual $i \in {\chi_E}^{-1}$. The source of i, ψ_i, is then randomly chosen from χ_ψ. Note that in the case of fitting the full model, a candidate drawn from $P_e(\boldsymbol{\psi})$ is set to be ψ_i if it is also in χ_ψ (see Section 4.2.5). After initialising the transmission network and times of events (i.e., E_j and I_j), the transmitted sequences are initialised sequentially in the order of E_j according to the evolutionary model specified by Equation 4.2.

4.3 Simulation studies

In this section we perform inference of transmission dynamics based on epidemics simulated under conditions that reflect real-world scenarios, with the aim of assessing the performance of our inference framework in a range of circumstances.

We also investigate the effect of having partial genetic data in two different ways, from which the results should bear implications for future study designs. First, we investigate the effect of *sub-sampling of exposures*, when sequence samples are available for only a subset of exposures. Second, for both economic and computational efficiency considerations, we also investigate the effect of *partial genome sequencing*, when only a section of the genome is sequenced from each sample collected. Note that as the transmitted sequences are imputed, unsampled exposures can be naturally included and the effect of sub-sampling of exposures can be therefore studied. Specifically, along with the experiment for the scenario with full sampling whereby every exposure is sampled, we also consider scenarios where a sequence sample and the corresponding sampling time may have a fixed probability to be excluded from observed data. To enable a valid comparison, a scenario with a higher sampling percentage always includes the samples in a scenario with a smaller sampling percentage. Also note that when genetic data are not available (i.e., 0% of the exposures are sampled) only the epidemic model in Section 4.2.1 is fitted.

4.3.1 Inference for epidemics with multiple clusters

To test our algorithm we first apply it to analyse spatio-temporal multiple-cluster epidemics simulated in a population of size $N = 150$, whose locations are generated independently from a uniform distribution over a square region, between times $t = 0$ and $t = t_{max} = 60$ (days). We assume that the epidemic start at time 0 and evolve according to Equation (4.1). We firstly consider parameterisation with $\alpha = 0.0004$, $\beta = 8.0$, $K(d_{ij}, \kappa) = \exp(-0.02 d_{ij})$, and that the sojourn times in classes E and I follow $Gamma(10, 0.5)$ and $Weibull(2, 2)$ distributions respectively. Pathogen sequences of length $n = 8000$ are transmitted upon infection and evolve according to Equation (4.2) with $\mu_1 = 0.002$ (bases per day) and $\mu_2 = 0.0005$ (bases per day). Also, each base of the master sequence G_M is drawn uniformly from the set $\omega_N = \{A, C, G, U\}$ and we let $p = 0.01$. We also consider a second parameterisation with a significantly higher primary transmission rate (hence expectedly with a larger number of clusters) and higher mutation rates. In particular, we have $\alpha = 0.002$, $\beta = 8.0$, $\mu_1 = 0.003$, $\mu_2 = 0.001$ with other model parameters being the same as those used above. We consider simulations with these two sets of parameters which give rise

to a three-cluster epidemic and a six-cluster epidemic respectively. We consider the case of (random) partial genome sequencing where $n = 1000$ (see later in Section 4.3.2 for a comparison between partial genome sequencing and full genome sequencing). The observations $\boldsymbol{y}$ constitute only the sequences sampled from exposed individuals and the corresponding known sampling times, a bounded range of the times and the precise locations of transitions from E to I and the precise times and locations of transitions from I to R that occur during the observation period.

We demonstrate the feasibility of imputing the distribution of background sequences and hence allow inference of multiple-cluster transmission graphs. Specifically, we impute the master sequence G_M and the model parameter p along with the imputations of other model parameters and unobserved data.

Estimation of the transmission graph and other model quantities

The (overall) *coverage rate* of an imputed transmission graph is here defined as the proportion of infections for which the correct source is identified in the network. The posterior distribution of the coverage rate is therefore a useful indicator of how well the imputed networks match the true network. From Figure 4.3 we first notice that in the case with full sampling, the transmission graph is typically recovered with near-complete accuracy. It is also clear that the means of the posterior distributions of the coverage rate increases with the proportion of exposed individuals being sampled.

Figure 4.4 and Figure 4.5 show the posterior distributions of the model parameters corresponding to three- and six-cluster epidemics respectively. Traceplots for the 100% and 50% sampling are shown in Figure 4.6 to Figure 4.9. We notice that the parameter p (which governs the variation of the master sequence) can be accurately estimated. Figure 4.4(a) and Figure 4.5(a) show that in general the credible intervals of the epidemiological parameters become narrower when more genetic data become available. This trend appears to be most prominent for β and κ, which is not surprising given their roles in determining the transmission graph and the fact we have shown above that the transmission graph is more accurately estimated when genetic data are more readily available. Figure 4.4(b) and Figure 4.5(b) show similar but much less prominent trends for the genetic model parameters. Note that as the times of transitions from E to I are known within a bounded range (see also Section 4.2.6), we do not observe significant differences among the scenarios for parameters γ and η. When the proportion of sampling further reduces, the estimates of model parameters, especially for the mutation rates and model parameters of latent period distributions, become less robust (not shown).

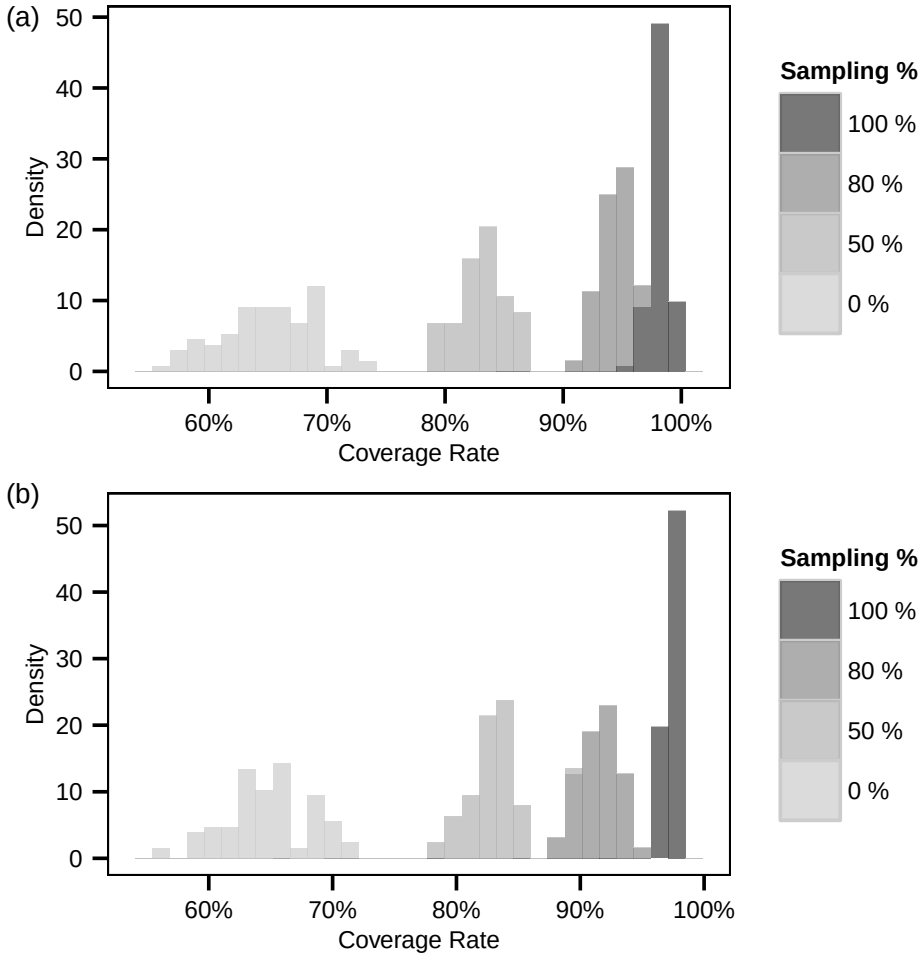


Figure 4.3: Posterior distributions of the overall coverage rate for the two multiple-cluster epidemics. (a) 3-cluster; (b) 6-cluster

Inference with a known latent period distribution We have so far considered estimations of the full set of model parameters under scenarios with at least 50% of the exposures are sampled. Figure 4.10 and Figure 4.11 demonstrate that a smaller proportion of sampling (e.g., 20%) may be tolerated if some of the model parameters, specifically the model parameters of the latent period distribution in this illustration, are assumed to be known.

Estimation of the master sequence G_M and the number of clusters Define ΔM as the number of bases differing between the imputed master sequence G_M and the actual one. Table 4.2 shows that the imputed master sequences match closely the correct sequence with an insignificantly decreasing similarity between two sequences when the proportion of sampling reduces in the three-cluster case. Table 4.3 shows that the number of clusters, N_c, is also well-recovered by the posterior samples with a slight tendency of over-estimation when the proportion of sampling reduces.

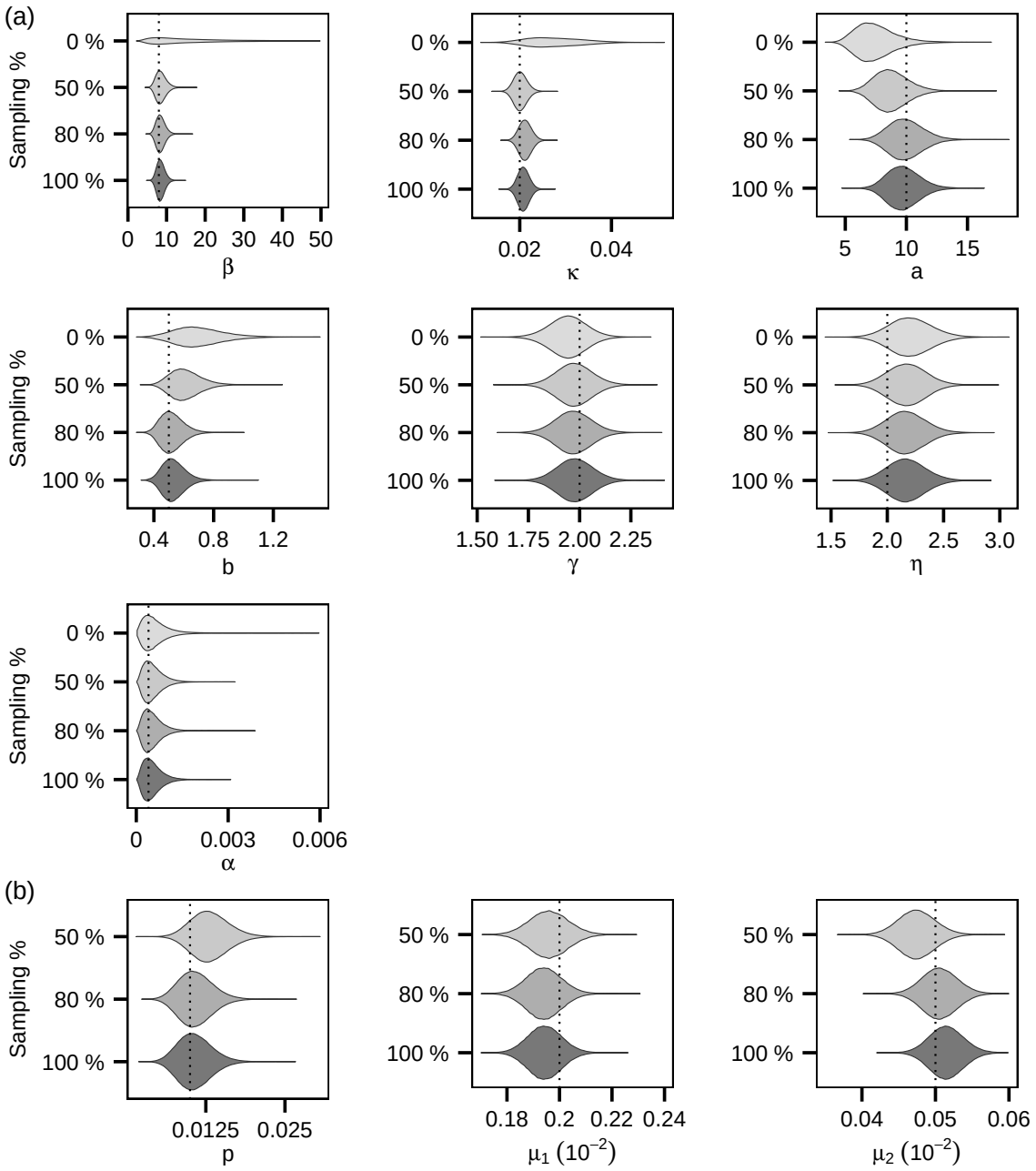


Figure 4.4: Posterior distributions of the model parameters (with the *three-cluster* epidemic). Dotted lines represent the true values of the model parameters. a) Epidemiological parameters; (b) Evolutionary model parameters

Table 4.2: Summaries of the posterior distribution of the number of differing bases between a particular imputed sequence and the actual master sequence G_M. The mean of ΔM is followed by the standard deviation in brackets

Sampling%	100%	80%	50%
ΔM (3-cluster)	0.28 (0.52)	0.43 (0.63)	0.64 (0.82)
ΔM (6-cluster)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)

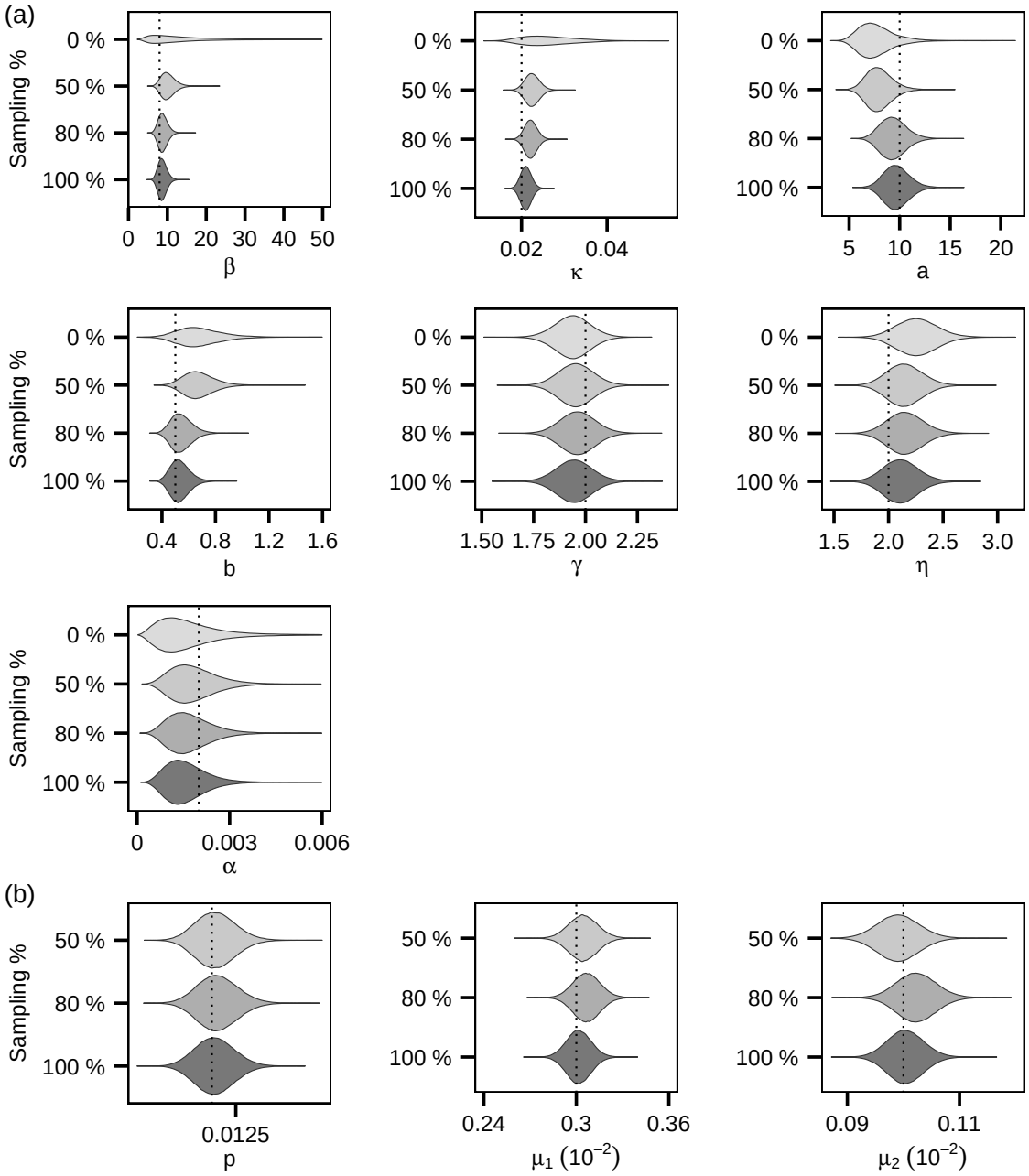


Figure 4.5: Posterior distributions of the model parameters (with the *six-cluster* epidemic). a) Epidemiological parameters; (b) Evolutionary model parameters

Table 4.3: Summaries of the posterior distribution of the number of cluster N_c. The mean of number of clusters is followed by the standard deviation in brackets

Sampling%	100%	80%	50%	0%
N_c (3-cluster)	3.04 (0.21)	3.08 (0.27)	3.13 (0.39)	3.73 (2.84)
N_c (6-cluster)	6.0 (0.0)	6.50 (0.70)	6.91 (1.02)	6.75 (5.06)

Estimation of individual coverage and identification of clusters

The (overall) coverage rate gives a broad measure of the recovery of the transmission graph. Here we examine the posterior distribution of the source of infection of

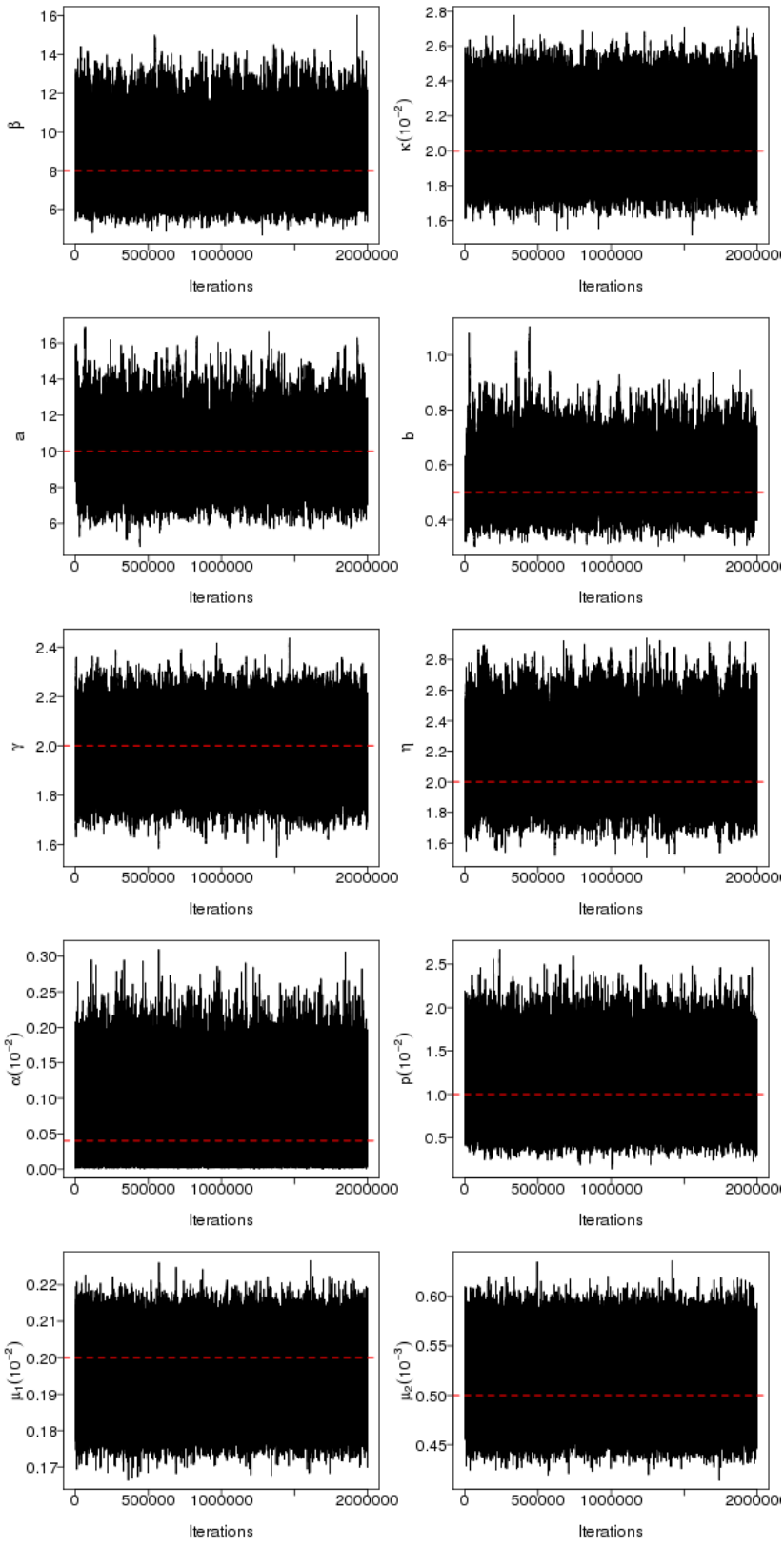


Figure 4.6: Traceplots of the posterior samples of model parameters in the case with 100% sampling (with the *three-cluster* epidemic). Dotted lines represent the true values of the model parameters.

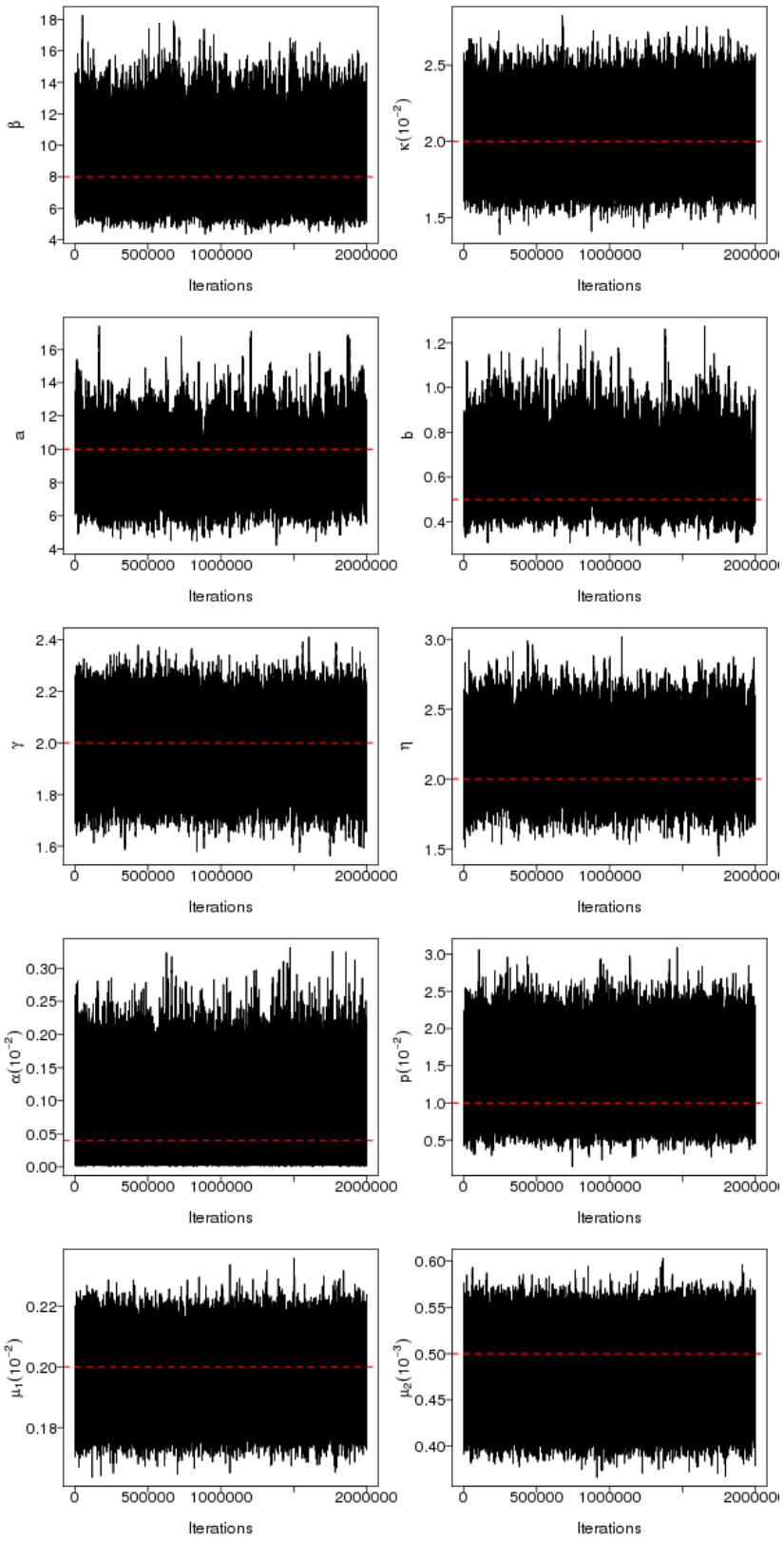


Figure 4.7: Traceplots of the posterior samples of model parameters in the case with 50% sampling (with the *three-cluster* epidemic). Dotted lines represent the true values of the model parameters.

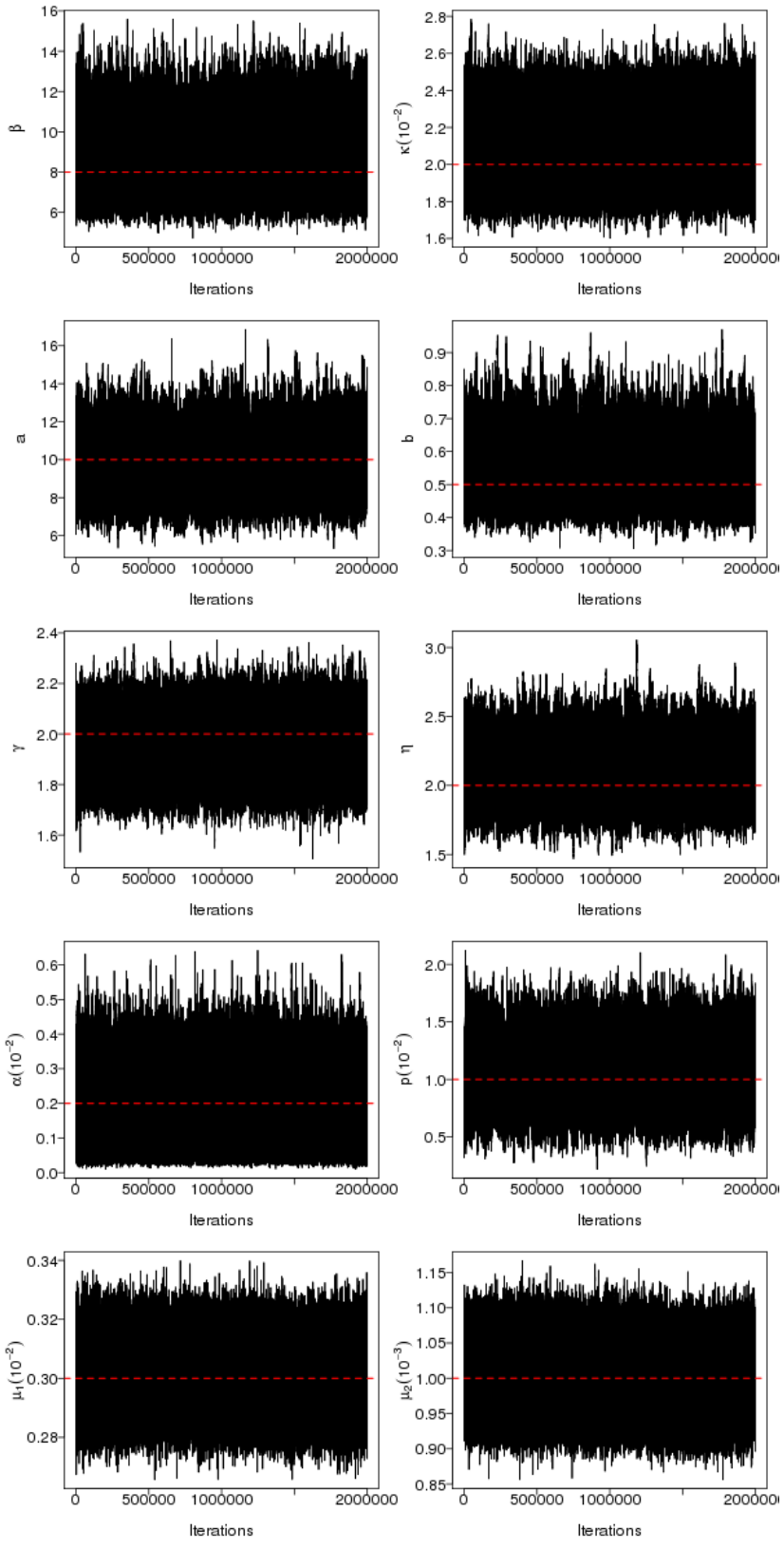


Figure 4.8: Traceplots of the posterior samples of model parameters in the case with 100% sampling (with the *six-cluster* epidemic). Dotted lines represent the true values of the model parameters.

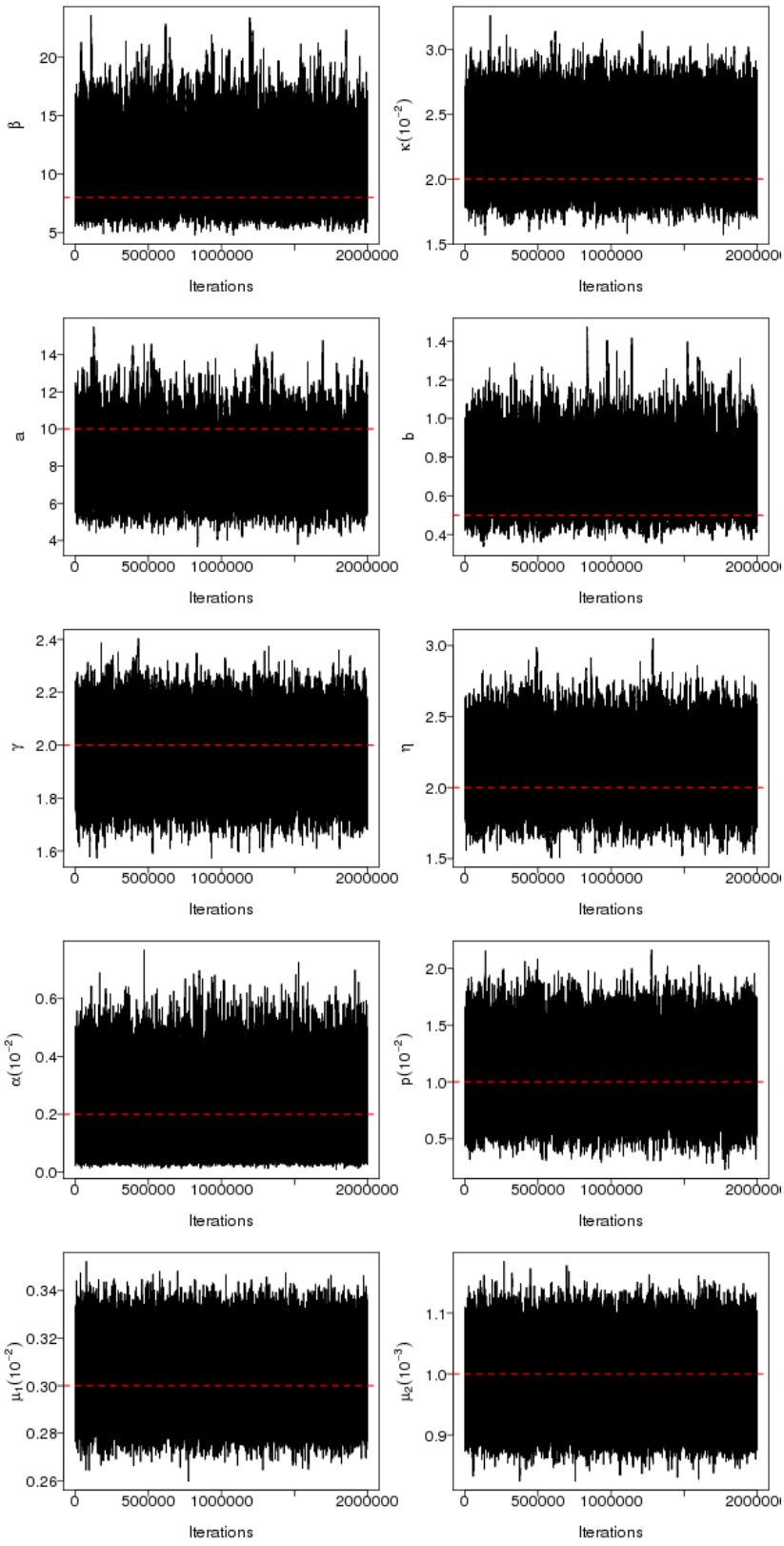


Figure 4.9: Traceplots of the posterior samples of model parameters in the case with 50% sampling (with the *six-cluster* epidemic). Dotted lines represent the true values of the model parameters.

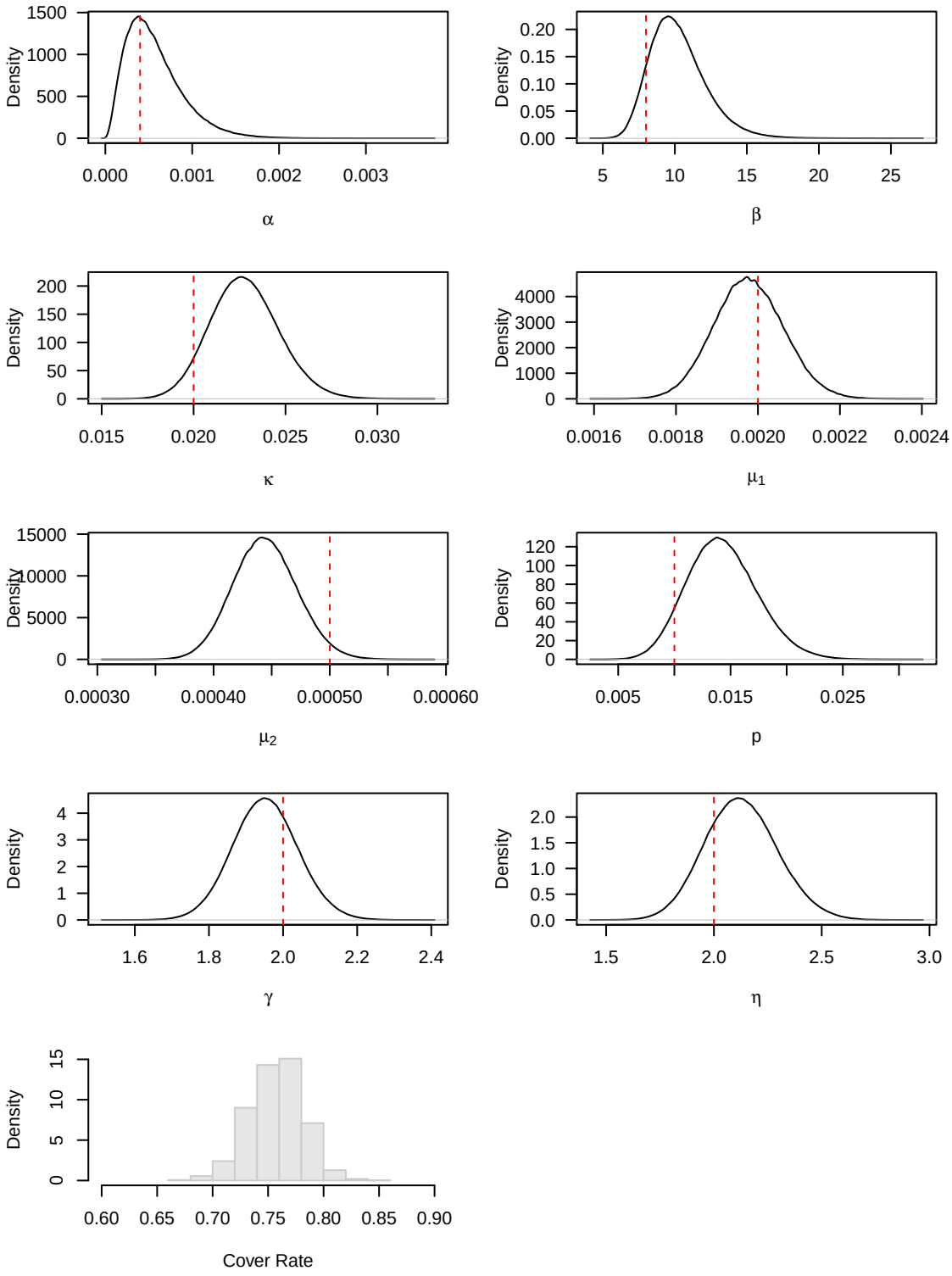


Figure 4.10: Posterior distributions of model parameters and the cover rate from fitting the three-cluster epidemic data with sampling proportion 20% (assuming the latent period distribution is known).

a particular exposure. Define the posterior *individual* coverage rate for a particular infection to be the proportion under the posterior distribution of transmission graph with which the true source of infection is correctly identified. Figure 4.12 shows the

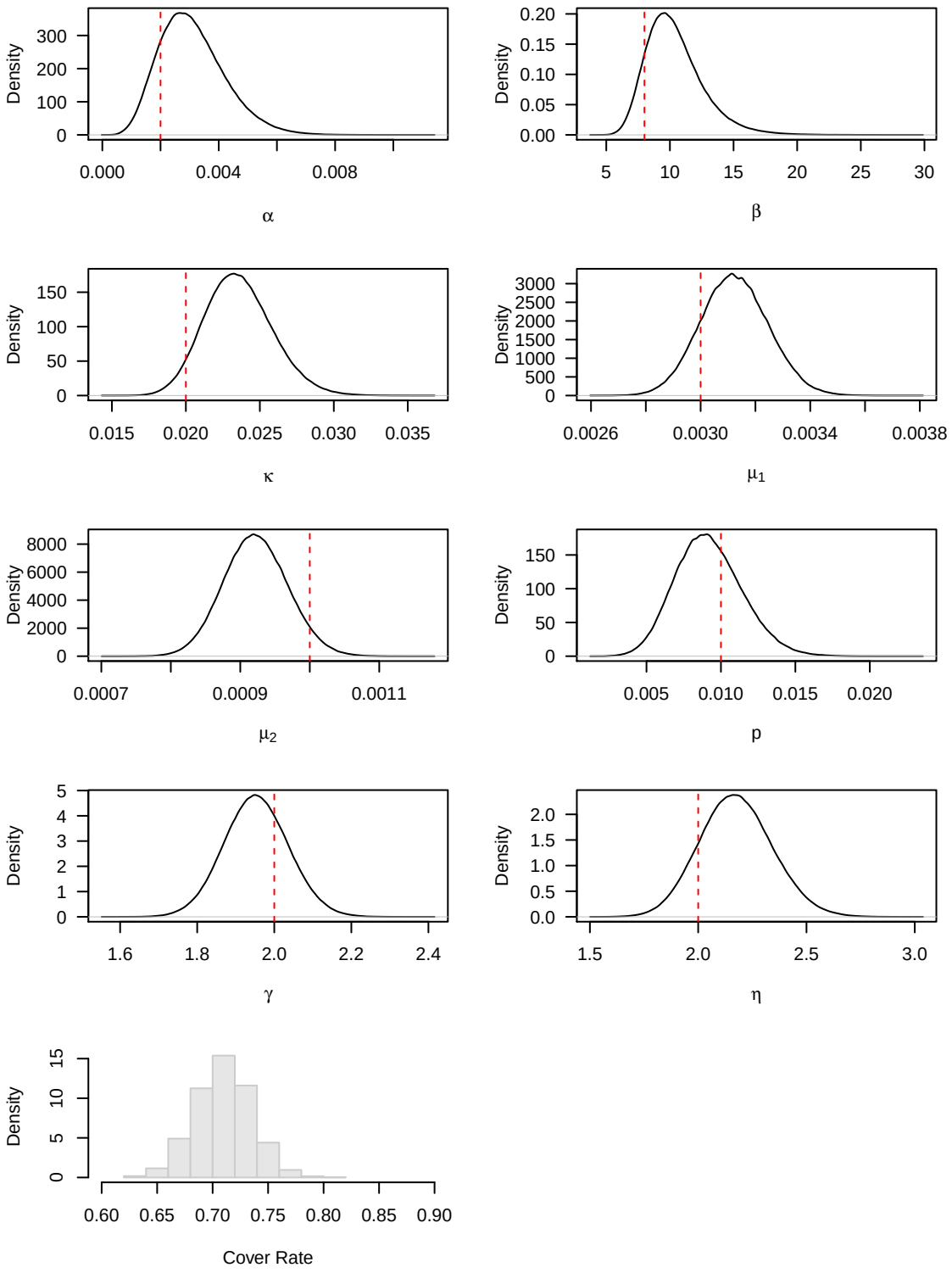


Figure 4.11: Posterior distributions of model parameters and the cover rate from fitting the six-cluster epidemic data with sampling proportion 20% (assuming the latent period distribution is known).

posterior individual coverage rate of all exposures at scenarios with different sampling percentages and it is noticed that the individual coverage rate in general increases with the sampling percentage. It is also noticed that the primary infections (indicated by

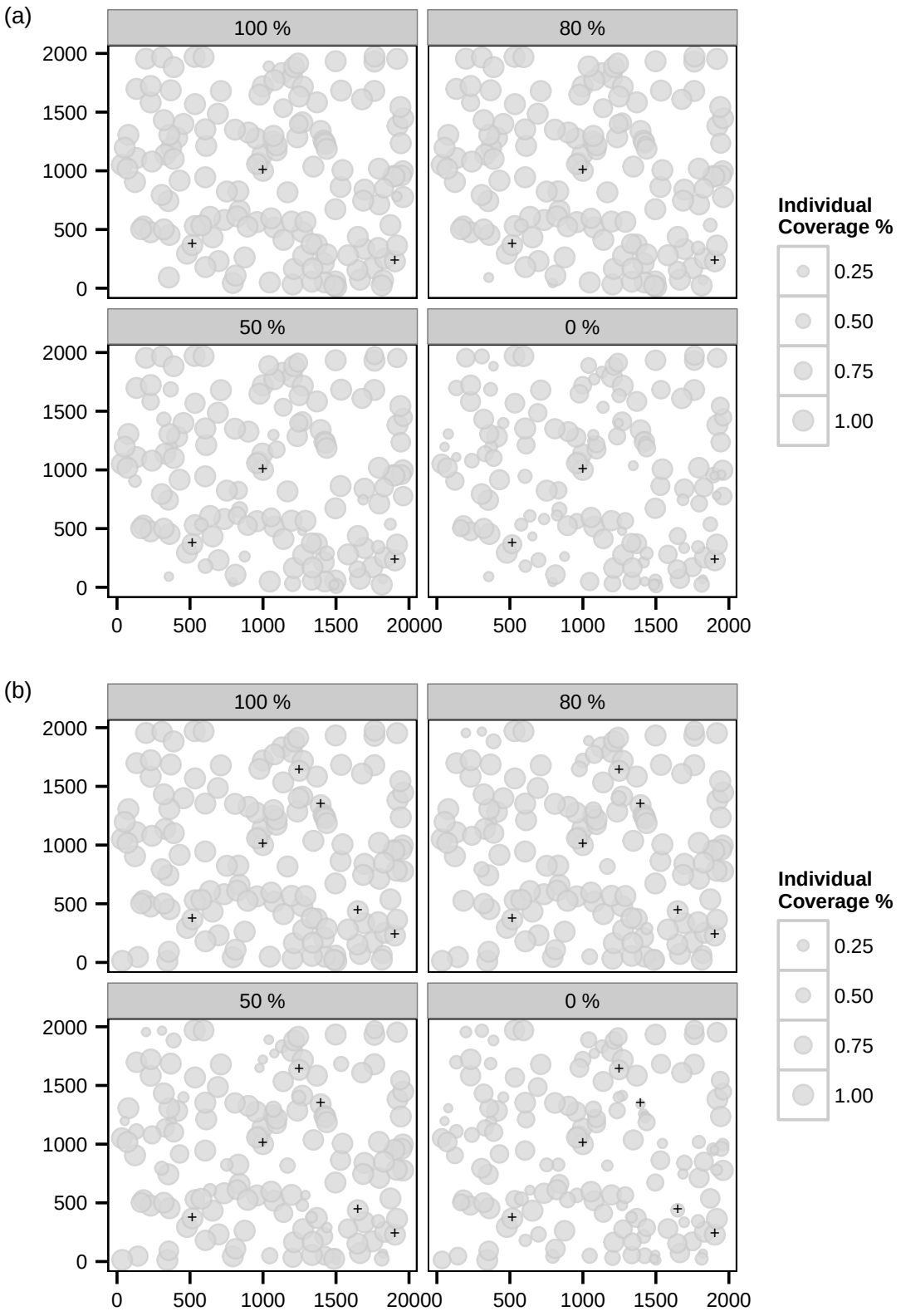


Figure 4.12: Posterior *individual coverage* of the source of infection (see main text) in scenarios with sampling 100%, 80%, 50% and 0%. The black + indicate the actual primary cases. (a) 3-cluster; (b) 6-cluster

black + signs) are frequently identified (i.e., high individual coverage rates), particularly in the scenarios with sequence samples.

Another natural question to ask is whether the clusters of transmission can be accurately identified by our analysis. In order to investigate this we consider two measures that can be calculated over posterior samples of the transmission graph and whose posterior expectations quantify the accuracy with which primary infections are identified in the inference. These are as follows:

1. For each infection we estimate the *cluster identification rate*, namely the proportion under the posterior distribution of the transmission graph with which the true primary infection leading to the given infection is correctly identified (i.e., the correct primary infection appears *as the root* of the sub-graph containing the given infection).

2. For each infection we estimate the *(primary) ancestor identification rate*, namely the proportion under the posterior distribution of the transmission graph with which the true primary infection leading to the given infection appears *on the path* from the infection to the root of the sub-graph.

Clearly, measure (1) will be lower than (2) since the conditions for 'success' are stronger. By estimating these quantities, are able to quantify the extent to which the link between primary and secondary infections, and hence the clusters of transmission, is accurately identified in the inferential procedure. For a given transmission graph, we can identify the total number of infections that are linked to the correct primary infection according to the criteria used in the definition of (1) and (2) above to provide two alternative summary statistics of the graph that capture the extent to which attribution to primary infection has been inferred in the graph.

Here we focus on the analysis of the six-cluster epidemic. From Figure 4.13 and Figure 4.14 we first notice that the primary-to-secondary infection links, and hence the clusters, can be reasonably inferred in the scenarios with sequence samples. Also, the difference between high and low sampling levels is insignificant compared to the difference of individual coverage rates observed in Figure 4.3(b) and to the difference of overall coverage rates observed in Figure 4.3. These results indicate that the clusters may be accurately identified even in scenarios with a relatively small percentage of sampling while the transmission graph may be less accurately inferred. Note that in the scenario with no sequence data the cluster identification rate for cluster 5 is low (see Figure 4.13), which indicates that the root of the cluster is not frequently identified as a primary infection (also see Figure 4.3(b)); nevertheless, the ancestors of the cases in this cluster can be accurately estimated (see Figure 4.14).

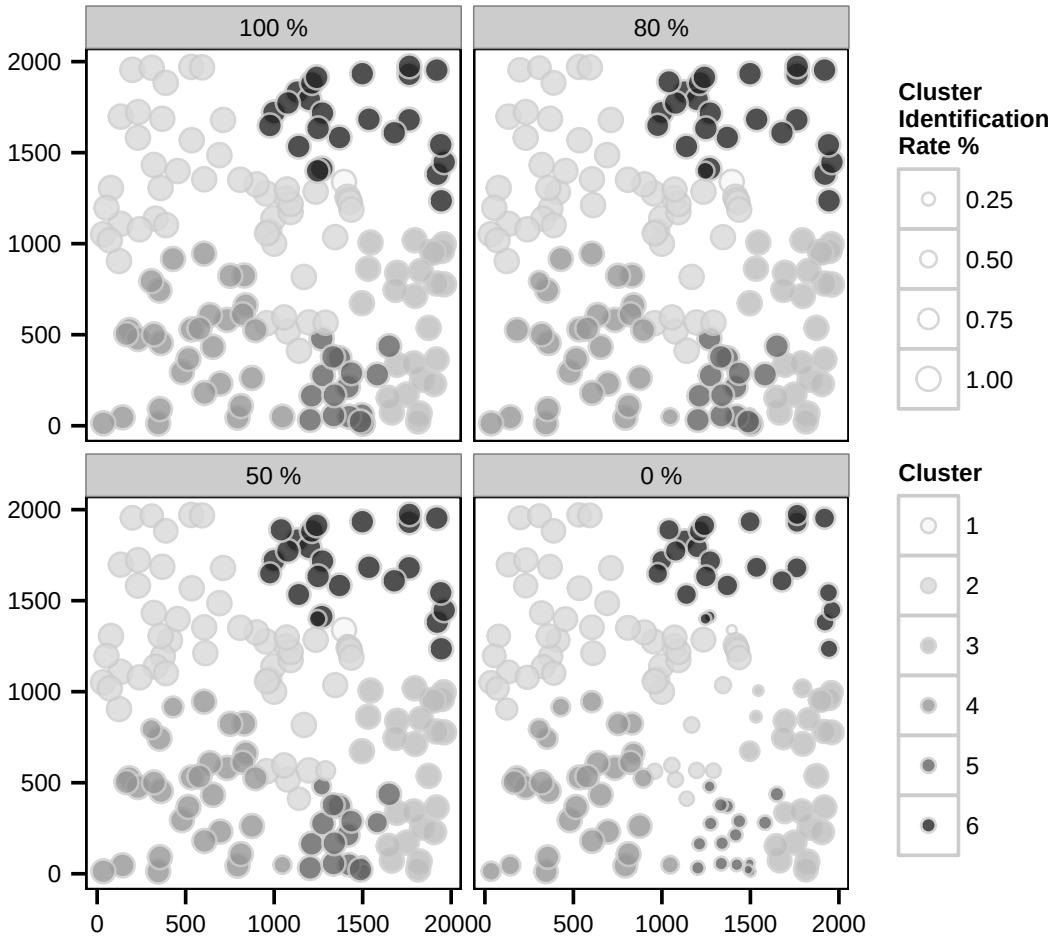


Figure 4.13: Posterior *cluster identification rate* of the infections (see main text), within each actual cluster of the *six-cluster* epidemic, in scenarios with sampling 100%, 80%, 50% and 0%.

Effect of mutation rates on inference

In previous sections we have chosen the model parameters for simulated scenarios, using values arising from practical considerations (Lau et al, 2014b; Morelli et al, 2012; Ypma et al, 2012; Ferguson et al, 2001). In particular, the number of nucleotide bases and the mutation rates are chosen to lie within the respective ranges of these quantities for common animal viruses (Mettenleiter and Sobrino, 2008; Morelli et al, 2012; Ypma et al, 2012). In this section, we also investigate the effect of mutation rates, parameters that make a key contribution to the likelihood, on inference by considering pathogens with much smaller mutations rates (e.g., foot-and-mouth disease pathogen) than those we have considered. Notably, results show that the estimation of the full set of model parameters is still feasible under the scenario with only 10% of sub-sampling (Figure 4.15), in contrast to 50% in previous sections. This could indicate that when mutation rates are higher, in which case the transmitted sequences on exposures may be more distinct individually to each other, more individual data (i.e., the sequence

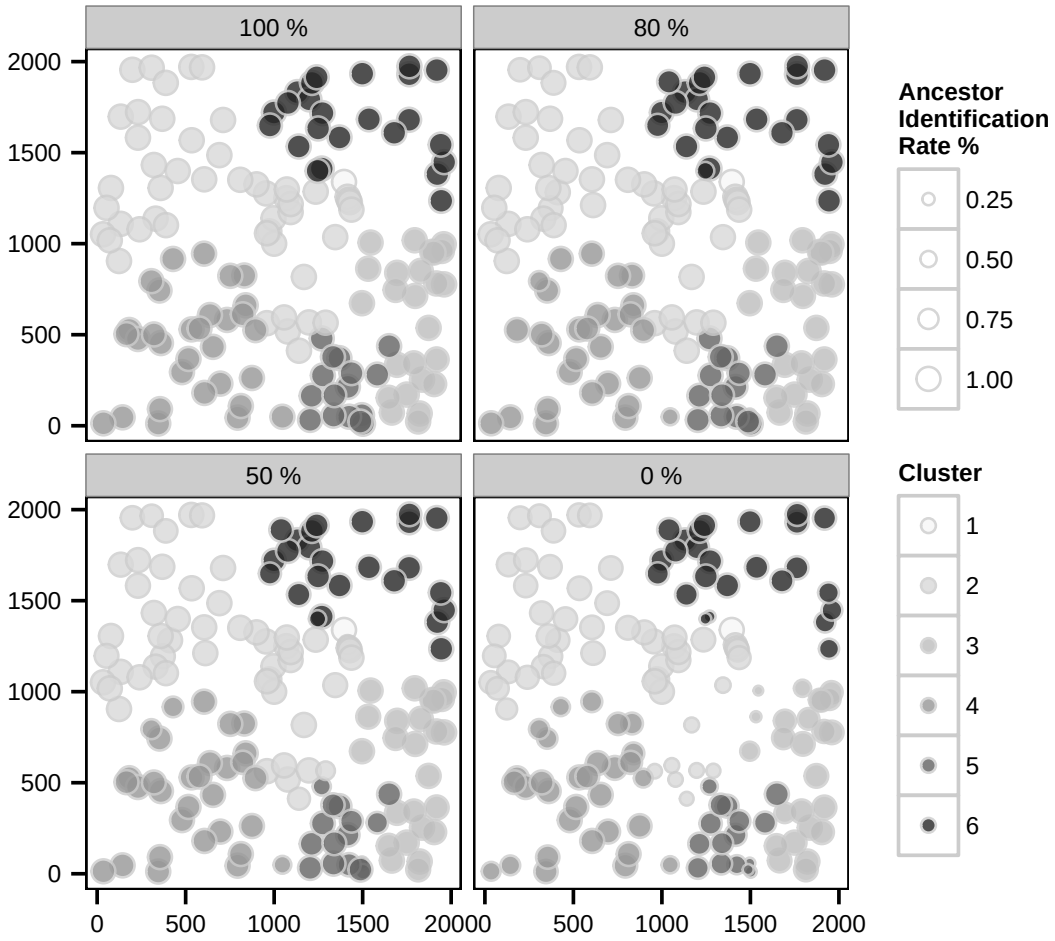


Figure 4.14: Posterior (primary) *ancestor identification rate* of the infections (see main text), within each actual cluster of the *six-cluster* epidemic, in scenarios with sampling 100%, 80%, 50% and 0%.

samples) on the exposures may be required for robust inference.

The mutation rates of foot-and-mouth disease (FMD) pathogens are much lower than that considered in the Section 4.3.1 and 4.3.2 but are known to be sufficiently high that the sampled sequences can provide significant information on the transmission graph (Ypma et al, 2012; Morelli et al, 2012). In this section we consider an epidemic with mutation rates in keeping with the FMD scenario. In particular, we set $\beta = 8.0$, $\mu_1 = 10^{-4}$, $\mu_2 = 5 \times 10^{-5}$ with other model parameters being set to the values used for simulating the three-cluster epidemic in Section 4.3.1. We note that the value of the variation parameter $p = 0.01$ may relate to the number of variant nucleotides among the sample sequences collected during the FMD outbreak in 2001 in UK (Cottam et al, 2008). In order to discern any resulting differences due to the change of mutation rates and genetic data, we consider a particular simulation yielding the same epidemic data as the three-cluster epidemic mentioned above.

Notably, Figure 4.15 shows that the estimation of the full set of model parameters is

still feasible under the scenario with only 10% of sub-sampling (in contrast to 50% in other sections with higher mutation rates). Results concerning the transmission graph and coverage rates are similar to previous sections (see Figure 4.16, Figure 4.17, Figure 4.18, Figure 4.19).

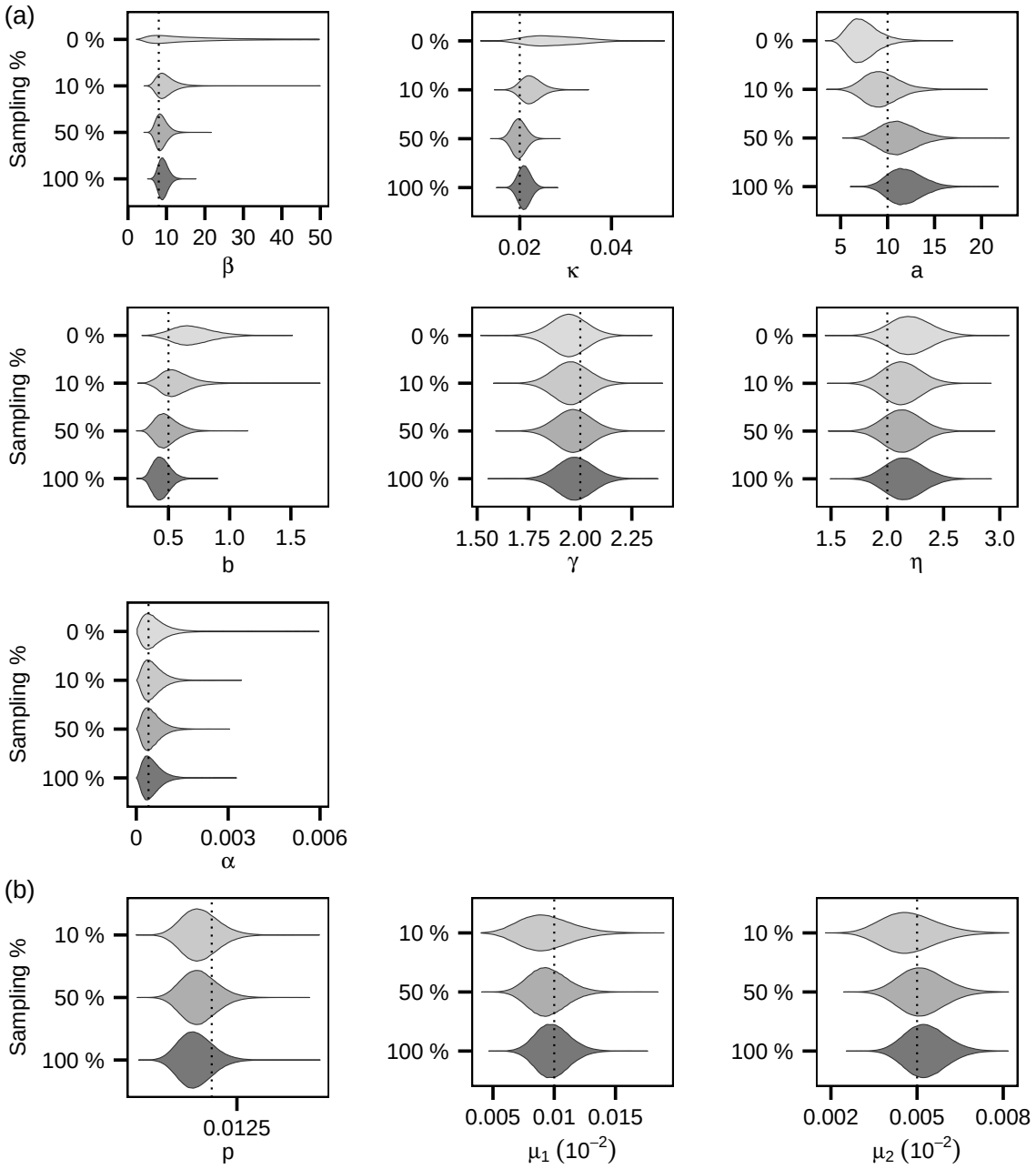


Figure 4.15: Posterior distributions of the model parameters for the epidemic with lower mutation rates.

Inference based on pseudo-likelihood approach

In this section we demonstrate the performance of the pseudo-likelihood approach proposed in Morelli et al (2012) (see Equation 4.8). We fit the same model considered in

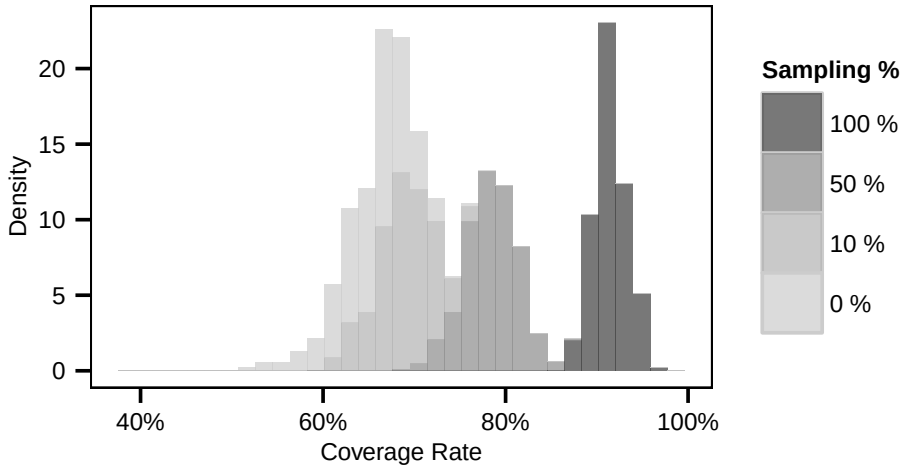


Figure 4.16: Posterior distributions of the overall coverage rate for the epidemic with lower mutation rates. Notice that, at the low sampling percentage, 10%, the availability of genetic data may not improve significantly the estimation of the coverage rates compared to the scenario without any samples.

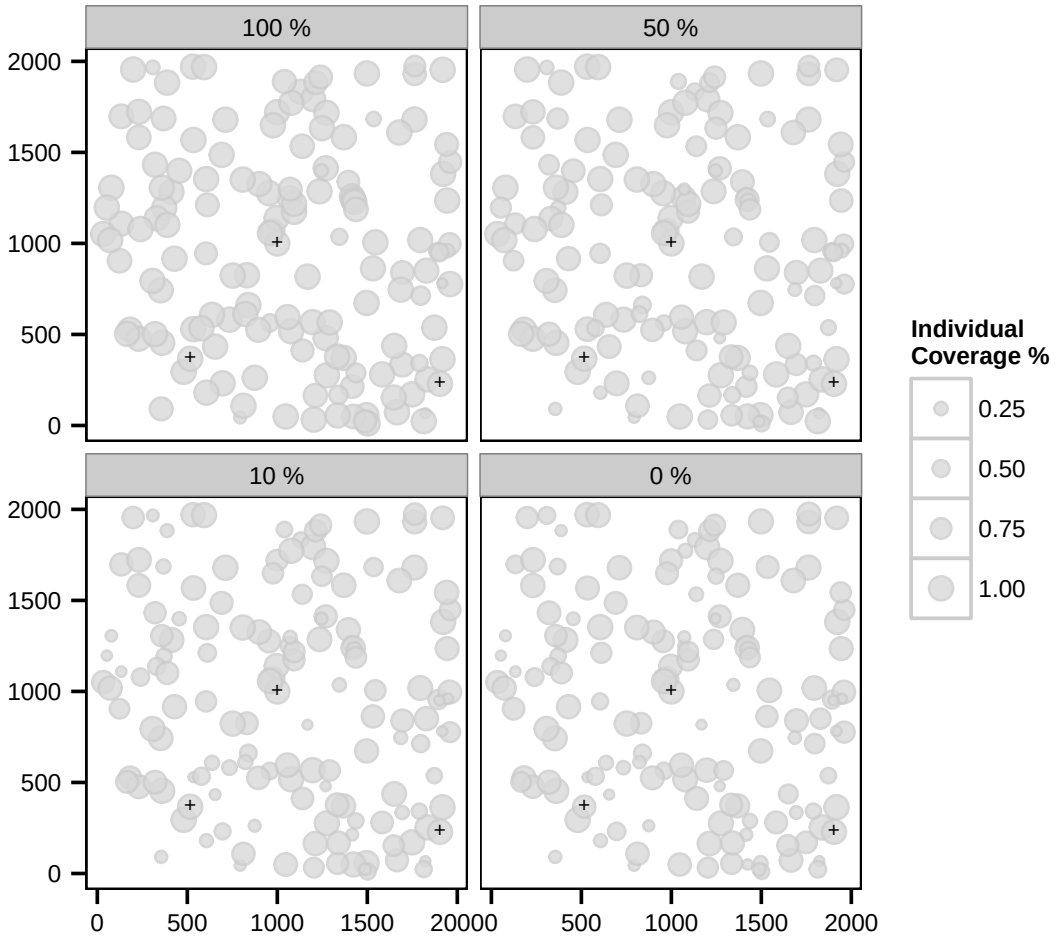


Figure 4.17: Posterior individual coverage of the sources of infection for the epidemic with lower mutation rates in scenarios with sampling 100%, 50%, 10% and 0%. The black + indicate the actual primary cases.

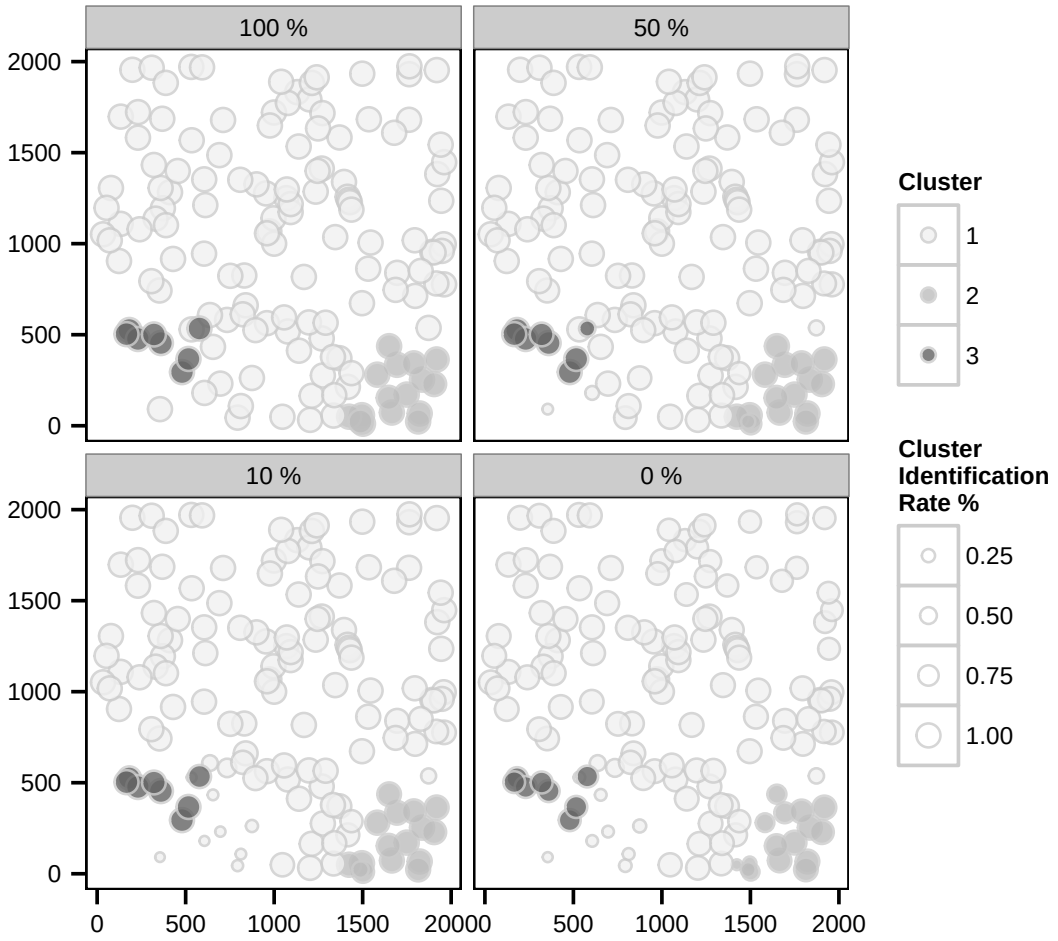


Figure 4.18: Posterior cluster identification rate of the infections, within each actual cluster of the epidemic with lower mutation rates, in scenarios with sampling 100%, 50%, 10% and 0%.

Section 4.3.1 to the 3-cluster epidemic but using the pseudo-likelihood to approximate the evolutionary process and not imputing the transmitted sequences. We consider only the full-sampling case where every exposure has a sampled sequence. Figure 4.20 shows that the posterior distributions of most of the model parameters deviate significantly from the true values (also compare with Figure 4.4 where our approach is deployed). Note that our approach can estimate the mutation rates reliably but we have to assume these rates to be known when using the pseudo-likelihood approach, for otherwise we are not able to obtain a converged and well-mixing chain (not shown here).

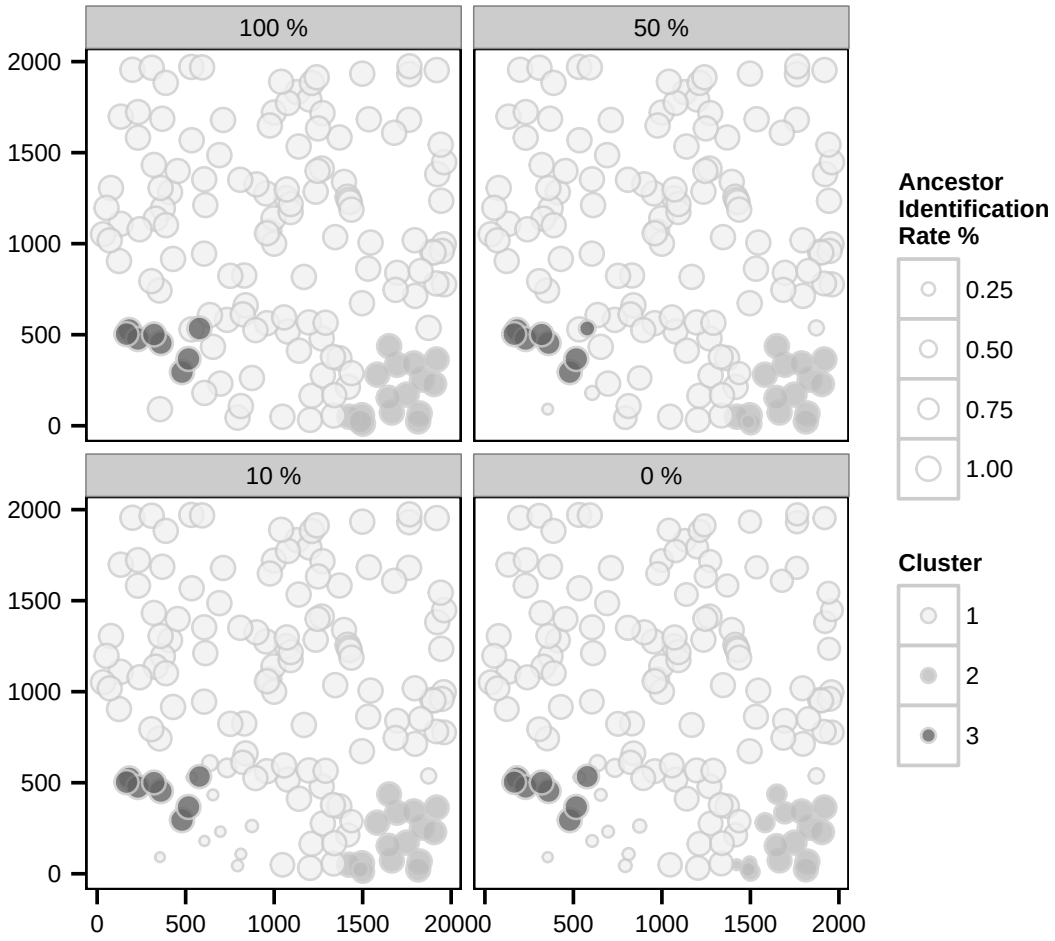


Figure 4.19: Posterior (primary) ancestor identification rate of the infections, within each actual cluster of the epidemic with lower mutation rates, in scenarios with sampling 100%, 50%, 10% and 0%.

4.3.2 Inference for epidemics with single cluster

In this section we consider the more restrictive single-cluster scenario, which is most commonly assumed in the literature. Our framework can be easily adapted to a single-cluster scenario by disregarding the process of generating the background sequences (see also Section 4.2.4). We assume $\alpha = 0.0004$, $\beta = 10.0$ and other model parameters being the same as those used for simulating the three-cluster epidemic (see also Section 4.3.1). We consider a particular simulation giving rise to a single-cluster epidemic. We also compare the case of *partial genome sequencing* with the case that where full genome sequencing is considered.

The transmission graph and the model parameters can be accurately estimated and the effect of sub-sampling of exposures is similar to that observed in multiple-cluster scenarios (see Figure 4.22 to 4.24). Figure 4.21 demonstrates that a higher degree of sequencing of the genome gives rise to narrower credible intervals for μ_1 and μ_2

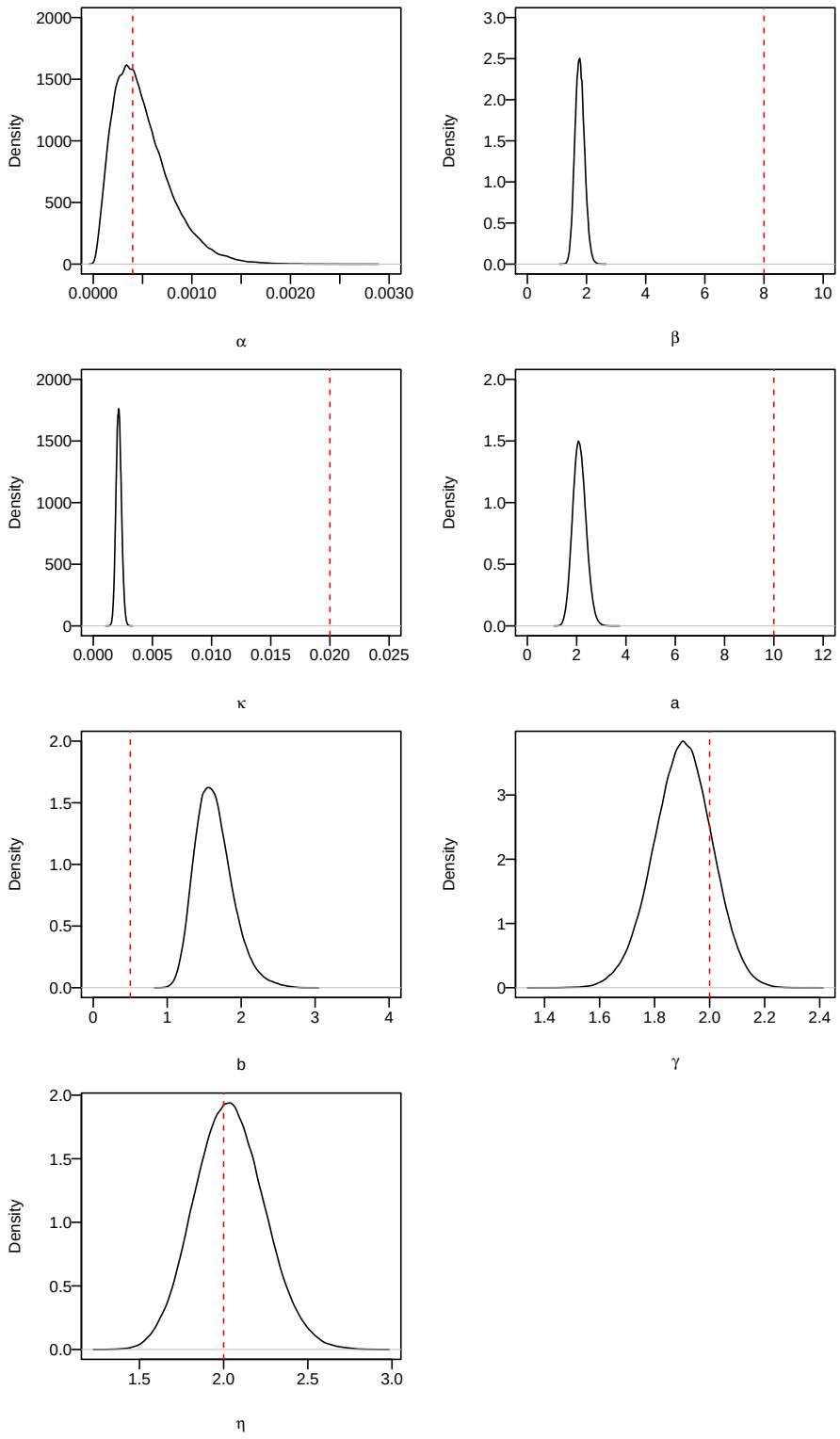


Figure 4.20: Posterior distributions of model parameters using pseudo-likelihood approach proposed in Morelli et al (2012), assuming mutation rates are known. Dotted lines indicate the true model parameters. Note that the model parameter p and the master sequence G_M are irrelevant in using the pseudo-likelihood approach.

compared to the case with partial genome sequencing. It reveals that partial genome sequencing may be sufficient if the transmission graph and epidemiological model

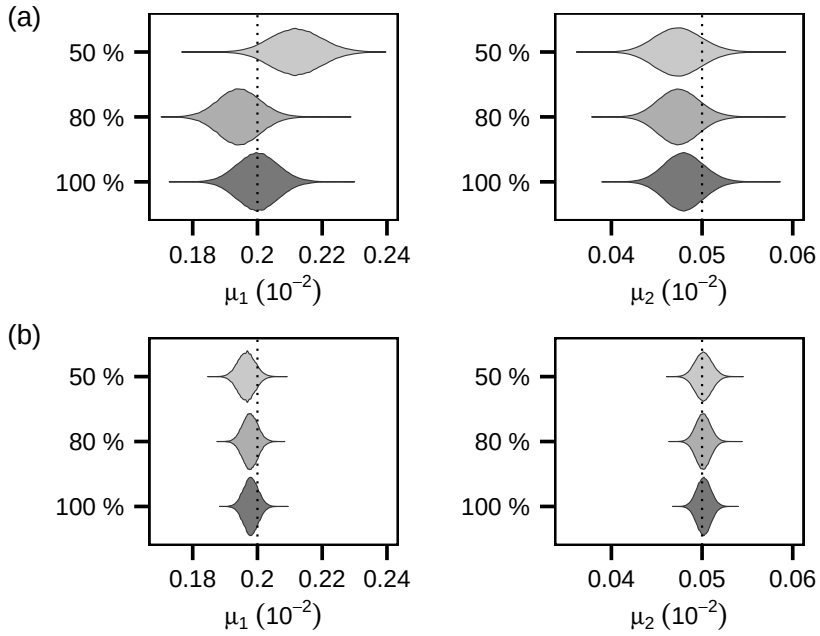


Figure 4.21: Posterior distributions of the mutation rates (with the single-cluster epidemic). (a) $n = 1000$; (b) $n = 8000$

parameters are of primary interest as the quality of the estimation appears robust to reduction of the amount of sequencing of the genome (Figure 4.22 to 4.24).

To show that the increasing genetic data systematically provides extra information on the transmission dynamics, extensive simulation studies are conducted in Section 4.4.

Computing time and other benchmarks

Our analysis was coded in C++ language (executed on a system with an Intel(R), i7-2600, 3.40GHz CPU). To provide a benchmark, we report the computing time and some key features of the Markov chain from the simulated example in Section 4.3.2 where full genome sequencing and full sampling of exposures were considered (i.e., population size $N = 150$, sequence length $n = 8000$ and sampling%=100). Convergence and mixing of the chain were assessed on the basis of visual inspection. The effective sample size of model parameters were computed using a package (Plummer et al, 2006) available in the statistical software R.

We obtained a converged and well-mixed chain with a reasonable effective sample size (400,000 iterations after 50,000 burn-in). The computing time was 63803.28 seconds (17.7 hours) which is considered to be practical and efficient (Ster et al, 2009; Morelli et al, 2012). Note that the computing time will be greatly reduced in the case of partial genome sequencing (e.g., when $n = 1000$). The effective size was

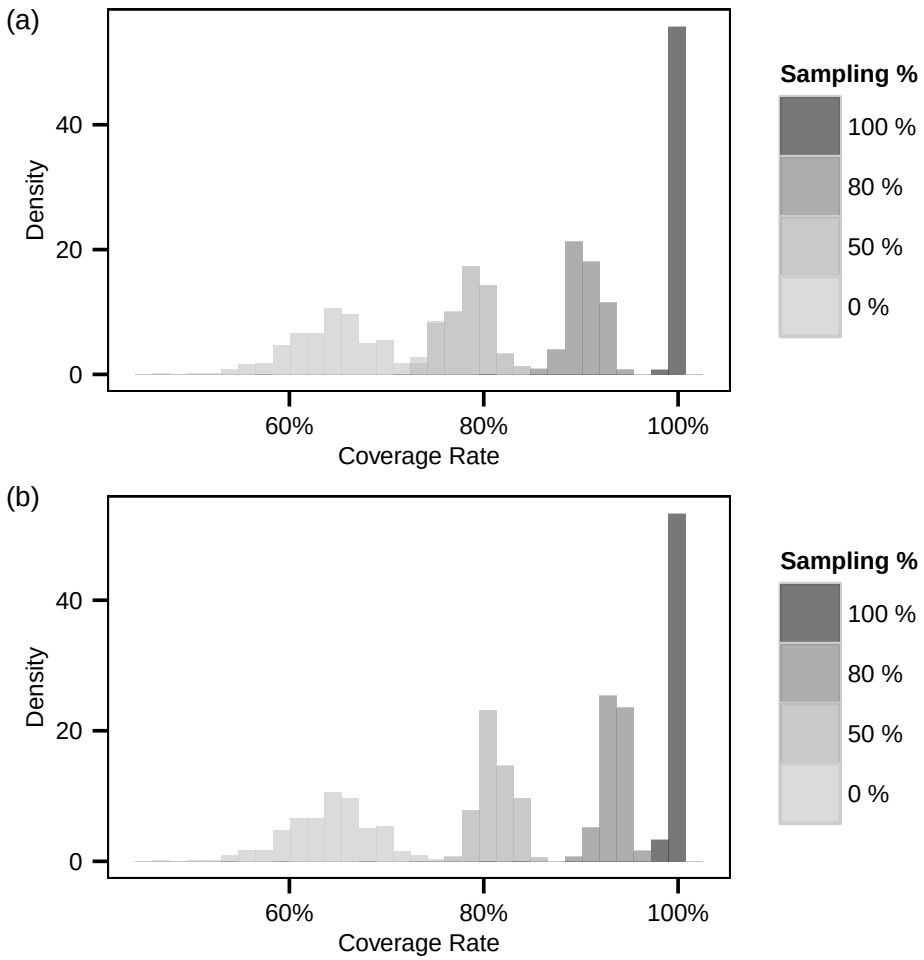


Figure 4.22: Posterior distributions of the overall coverage rate (with the single-cluster epidemic). (a) $n = 1000$; (b) $n = 8000$

$Eff_{\boldsymbol{\theta}} = (286, 912, 998, 5380, 1950, 7416, 9945, 30133)$ with elements corresponding to parameters in $\boldsymbol{\theta} = (\beta, a, b, \gamma, \eta, \kappa, \mu_1, \mu_2)$.

4.4 More simulated epidemics

In this section we consider 15 random independent replicates of epidemics in which there are respectively 5 of them simulated from each of the 3 sets of the model parameters adopted in the Section 4.3.1 where multiple cluster scenario was investigated. All the epidemics considered here are of more than one cluster. To recap, compared to the *first set* of model parameters, the *second set* of model parameters is characterised by a higher background transmission rate and hence is expected to give rise to epidemics with higher number of clusters than those from the first and third set of model parameters; the *third set* of model parameters is characterised by the lower mutation rates which match with foot-and-mouth disease.

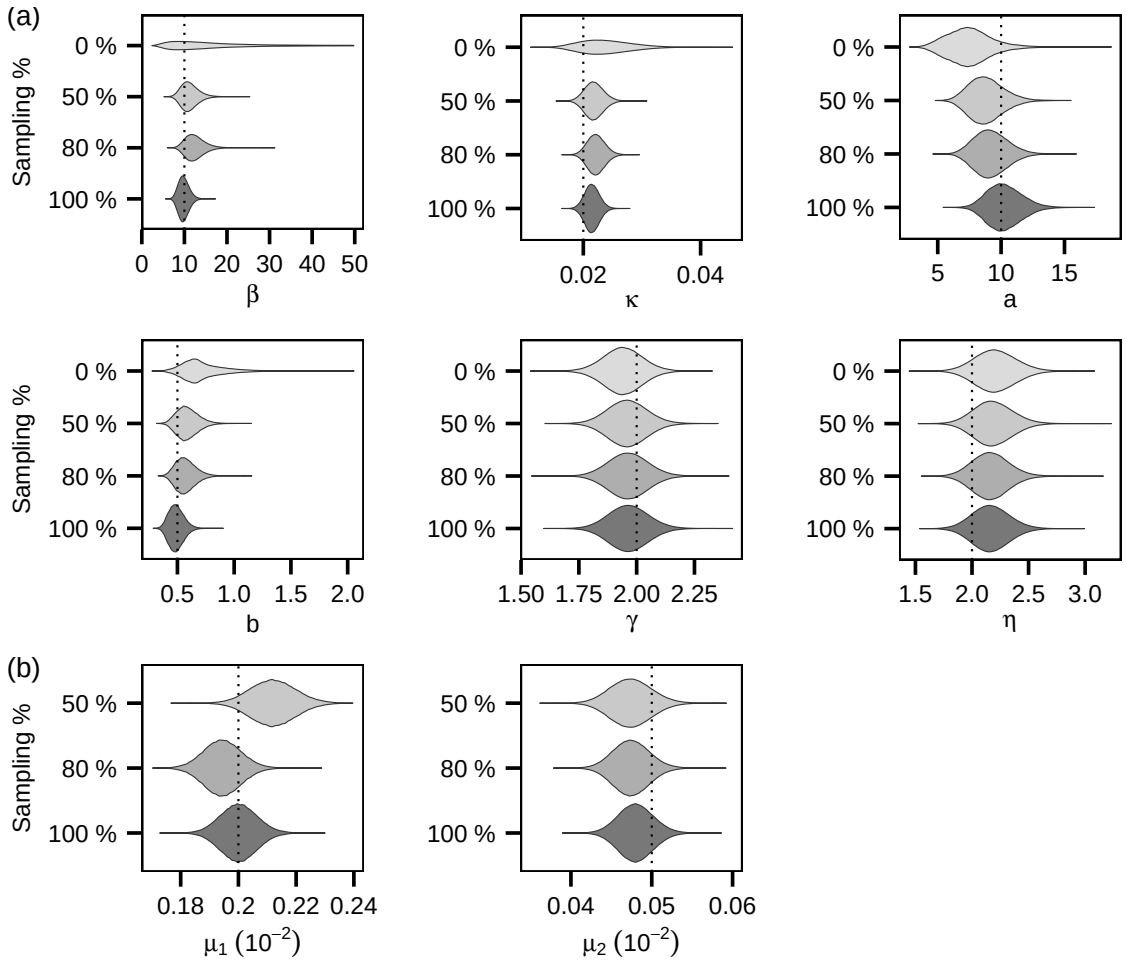


Figure 4.23: Violin plots showing the posterior distributions of the model parameters (with the single-cluster epidemic and number of bases $n = 1000$). Dashed lines represent the actual values of the model parameters. (a) Epidemiological parameters; (b) Evolutionary model parameters

For epidemics simulated from the first and second sets of model parameters, we compare the estimation performance at sampling levels 100%, 50% and 0% of the exposures. For epidemics simulated from the third set of model parameters, we compare the estimation performance at sampling levels 100%, 10% and 0% of the exposures.

Table 4.4 to 4.6 show the absolute difference between the number of clusters obtained from the posterior samples and the actual number of clusters, denoted as ΔN_c. They also show the number of different bases (out of 1,000) between the imputed master sequence G_M and the actual ones, denoted as ΔM, and the overall coverage rate obtained from the posterior samples. It is observed that ΔN_c and ΔM in general increase when the sampling percentage reduces. When no genetic data are available the mean number of ΔN_c and its variation are quite significant. Also, comparing the values of ΔM from Table 4.4 and Table 4.5 reveals that the estimation of G_M may become more reliable when the number of cluster increases. In fact, it is observed that

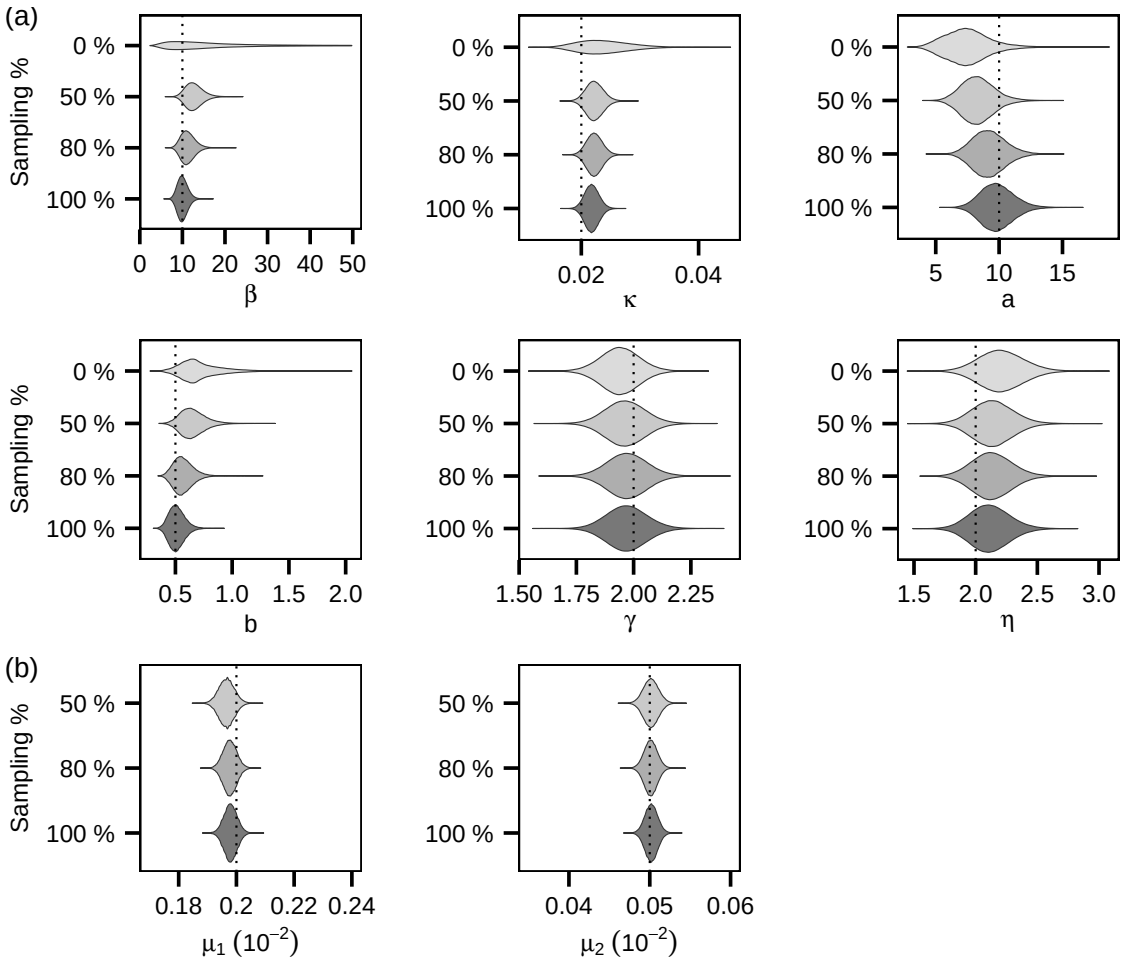


Figure 4.24: Posterior distributions of the model parameters (with the single-cluster epidemic and number of bases $n = 8000$)

when there are less than 3 clusters in the actual epidemic (e.g., Replicate 1, Replicate 3 and Replicate 5 in Table 4.4), ΔM becomes more significant. The coverage rate increases with the sampling percentage and becomes less dispersed.

In previous sections, it is observed that dispersion and hence the uncertainties in the model parameters estimates in general increase when the sampling % drops and this effect appears to be most dominant for the secondary transmission rate β and the spatial kernel parameter κ. Table 4.7 to Table 4.9, which show the sample means and standard deviations of the posterior samples of β and κ, suggest similar findings. Note that for the third set (characterised by lower mutation rates and shown to have higher tolerance to level of sub-sampling) we have considered sampling level 10% (instead of 50%) and the difference with 0% is not very significant.

Table 4.4: Values of ΔN_c, ΔM and the coverage rate obtained from the posterior samples from fitting 5 independent replicates of multiple-cluster epidemics simulated from the *first set* of model parameters with $\alpha = 0.0004$, $\beta = 8$, $\kappa = 0.02$, $a = 10$, $b = 0.5$, $\gamma = 2.0$, $\eta = 2.0$, $p = 0.01$, $\mu_1 = 0.002$ and $\mu_2 = 0.0005$. The mean values are followed by the standard deviations in the bracket.

	ΔN_c			ΔM		*Coverage*(%)		
Sampling%	100%	50%	0%	100%	50%	100%	50%	0%
Replicate 1	0.01 (0.1)	0.18 (0.44)	1.60 (3.01)	19.3 (3.32)	19.0 (3.10)	99.9 (0.28)	83 (1.9)	58.5 (3.3)
Replicate 2	0.0 (0.05)	0.35 (0.61)	2.60 (4.59)	0.26 (0.51)	1.0 (0.84)	98.8 (0.31)	87.7 (1.8)	65.7 (4.6)
Replicate 3	0.04 (0.19)	0.22 (0.47)	0.95 (3.04)	24.7 (2.53)	29.37 (2.6)	99.7 (0.39)	85 (2.2)	62.7 (4.3)
Replicate 4	0.07 (0.25)	0.41 (0.62)	1.59 (3.27)	0.03 (0.17)	0.09 (0.28)	98.6 (0.50)	83.3 (2.2)	62.8 (4.1)
Replicate 5	0.0 (0.07)	0.25 (0.52)	1.03 (2.58)	14.3 (2.8)	13.1 (2.6)	99.9 (0.17)	82.8 (2.5)	61.4 (4.1)

Table 4.5: Values of ΔN_c, ΔM and the coverage rate obtained from the posterior samples from fitting 5 independent replicates of multiple-cluster epidemics simulated from the *second set* of model parameters, which is characterised by a higher background transmission rate and higher mutation rates compared to the first set of model parameters, with $\alpha = 0.002$, $\beta = 8.0$, $\kappa = 0.02$, $a = 10$, $b = 0.5$, $\gamma = 2.0$, $\eta = 2.0$, $p = 0.01$, $\mu_1 = 0.003$ and $\mu_2 = 0.001$. The mean values are followed by the standard deviations in the bracket. Note that ΔM are all zero and it dndicates that the number of cluster (six clusters in this scenario) is important for the inference of the master sequence.

	ΔN_c			ΔM		*Coverage*(%)		
Sampling%	100%	50%	0%	100%	50%	100%	50%	0%
Replicate 1	0.12 (0.34)	2.11 (1.29)	3.44 (4.14)	0.0 (0.0)	0.0 (0.0)	98.3 (0.57)	83.7 (2.2)	63.3 (4.5)
Replicate 2	0.01 (0.1)	0.50 (0.73)	2.87 (4.58)	0.0 (0.0)	0.0 (0.0)	99.7 (0.36)	85.2 (2.1)	60.7 (4.3)
Replicate 3	0.14 (0.35)	1.18 (0.95)	2.43 (5.67)	0.0 (0.0)	0.0 (0.0)	99.1 (0.39)	75.5 (2.2)	61.6 (4.4)
Replicate 4	0.0 (0.0)	1.17 (1.09)	3.89 (4.94)	0.0 (0.0)	0.0 (0.0)	99.6 (0.4)	81.8 (2.3)	57.7 (4.7)
Replicate 5	0.0 (0.0)	2.20 (1.51)	2.77 (4.07)	0.0 (0.0)	0.0 (0.0)	99.2 (0.35)	79.8 (2.2)	65.2 (3.9)

Table 4.6: Values of ΔN_c, ΔM and the coverage rate obtained from the posterior samples from fitting 5 independent replicates of multiple-cluster epidemics simulated from the *third set* of model parameters, which is characterised by much lower mutation rates to match the foot-and-mouth disease compared to the first set of model parameters, with $\alpha = 0.0004$, $\beta = 8.0$, $\kappa = 0.02$, $a = 10$, $b = 0.5$, $\gamma = 2.0$, $\eta = 2.0$, $p = 0.01$, $\mu_1 = 10^{-4}$ and $\mu_2 = 5 \times 10^{-5}$. The mean values are followed by the standard deviations in the bracket.

	ΔN_c			ΔM		$Coverage(\%)$		
Sampling%	100%	10%	0%	100%	10%	100%	10%	0%
Replicate 1	0.01 (0.115)	0.61 (0.82)	1.54 (3.18)	12.06 (2.4)	12.0 (2.3)	89.4 (2.0)	60.3 (3.3)	58.5 (3.4)
Replicate 2	0.04 (0.20)	1.2 (1.15)	2.17 (3.58)	0.48 (0.50)	1.56 (0.66)	92.2 (1.7)	68 (3.0)	65.7 (3.7)
Replicate 3	0.02 (0.15)	0.49 (0.83)	0.88 (2.79)	16.93 (2.65)	18.38 (3.13)	93.6 (1.6)	66.4 (3.3)	62.6 (3.9)
Replicate 4	0.12 (0.35)	0.75 (0.56)	2.09 (2.59)	0.0 (0.0)	0.71 (0.71)	93.2 (1.6)	67.1 (3.3)	62.7 (4.0)
Replicate 5	0.0 (0.0)	0.32 (0.57)	0.74 (1.90)	4.89 (1.37)	5.71 (1.49)	94.9 (1.7)	68.1 (3.7)	64.5 (4.0)

Table 4.7: Sample means and standard deviations of the posterior samples of β and κ obtained from fitting 5 independent replicates of multiple-cluster epidemics simulated from the *first set* of model parameters with $\alpha = 0.0004$, $\beta = 8$, $\kappa = 0.02$, $a = 10$, $b = 0.5$, $\gamma = 2.0$, $\eta = 2.0$, $p = 0.01$, $\mu_1 = 0.002$ and $\mu_2 = 0.0005$.

	$\beta = 8.0$			$\kappa = 0.02$		
Sampling%	100%	50%	0%	100%	50%	0%
Replicate 1	8.39 (1.13)	7.38 (1.12)	10.08 (4.29)	0.018 (0.001)	0.018 (0.0012)	0.021 (0.0029)
Replicate 2	6.91 (0.82)	7.45 (1.02)	10.91 (8.73)	0.019 (0.0011)	0.019 (0.0012)	0.021 (0.0045)
Replicate 3	7.92 (1.01)	8.92 (1.34)	12.64 (6.71)	0.019 (0.0011)	0.019 (0.0012)	0.022 (0.0035)
Replicate 4	7.64 (1.02)	8.28 (1.29)	12.73 (7.95)	0.02 (0.0012)	0.021 (0.0014)	0.024 (0.0048)
Replicate 5	8.90 (1.27)	10.02 (1.96)	9.57 (4.78)	0.02 (0.0012)	0.021 (0.0015)	0.022 (0.0031)

Table 4.8: Sample means and standard deviations of the posterior samples of β and κ obtained from fitting 5 independent replicates of multiple-cluster epidemics simulated from the *second set* of model parameters with $\alpha = 0.002$, $\beta = 8.0$, $\kappa = 0.02$, $a = 10$, $b = 0.5$, $\gamma = 2.0$, $\eta = 2.0$, $p = 0.01$, $\mu_1 = 0.003$ and $\mu_2 = 0.001$.

	$\beta = 8.0$			$\kappa = 0.02$		
Sampling%	100%	50%	0%	100%	50%	0%
Replicate 1	8.29 (1.17)	8.21 (1.57)	6.21 (2.17)	0.021 (0.0012)	0.020 (0.0015)	0.018 (0.0023)
Replicate 2	8.13 (1.11)	7.89 (1.25)	8.47 (4.11)	0.021 (0.0013)	0.019 (0.0013)	0.020 (0.0033)
Replicate 3	8.16 (1.09)	10.58 (1.98)	14.91 (7.87)	0.019 (0.0012)	0.020 (0.0013)	0.023 (0.0036)
Replicate 4	8.60 (1.25)	8.88 (1.62)	8.75 (4.91)	0.021 (0.0012)	0.021 (0.0015)	0.020 (0.004)
Replicate 5	7.80 (1.06)	9.12 (1.45)	10.22 (5.83)	0.021 (0.0012)	0.021 (0.0013)	0.022 (0.0036)

Table 4.9: Sample means and standard deviations of the posterior samples of β and κ obtained from fitting 5 independent replicates of multiple-cluster epidemics simulated from the *third set* of model parameters with $\alpha = 0.0004$, $\beta = 8.0$, $\kappa = 0.02$, $a = 10$, $b = 0.5$, $\gamma = 2.0$, $\eta = 2.0$, $p = 0.01$, $\mu_1 = 10^{-4}$ and $\mu_2 = 5 \times 10^{-5}$.

	$\beta = 8.0$			$\kappa = 0.02$		
Sampling%	100%	10%	0%	100%	10%	0%
Replicate 1	9.79 (1.61)	7.36 (1.52)	10.04 (4.33)	0.019 (0.0013)	0.018 (0.0016)	0.021 (0.003)
Replicate 2	9.69 (1.55)	9.76 (2.59)	13.99 (8.9)	0.021 (0.0013)	0.022 (0.002)	0.024 (0.0043)
Replicate 3	8.59 (1.25)	8.22 (1.71)	12.60 (6.62)	0.019 (0.0012)	0.019 (0.0017)	0.022 (0.0035)
Replicate 4	9.02 (1.41)	8.92 (2.47)	12.73 (7.95)	0.022 (0.0014)	0.023 (0.0023)	0.024 (0.005)
Replicate 5	9.02 (1.36)	8.82 (1.95)	13.46 (7.99)	0.021 (0.0014)	0.021 (0.0018)	0.023 (0.0036)

4.5 Contribution of genetic data to model assessment

In Chapter 3 we have shown that effective model assessment of a general spatio-temporal model may be achieved by proposing suitably designed non-centred parameterisation schemes and imputing the corresponding residuals, whose sampling distributions are known, in such a manner that posterior distributions are sensitive to mis-specifications of particular components of the model. In this section, we investigate how the genetic data may help in assessing, in particular, the goodness-of-fit of a specified *spatial kernel* by utilising the so-called *Infection-link Residual (ILR)*. Specifically, we consider fitting three forms of spatial kernel:

- An exponentially-bounded kernel (Kernel A): $K(d_{ij}, \kappa_1) = exp(-\kappa_1 d_{ij})$;
- A power-law kernel (Kernel B): $K(d_{ij}, \kappa_2) = d_{ij}^{-\kappa_2}$;
- A Cauchy-type kernel (Kernel C): $K(d_{ij}, \kappa_3) = \frac{1}{\kappa_3\{1+(\frac{d_{ij}}{\kappa_3})^3\}}$.

It is noted that Kernel A is the correct spatial kernel.

To recap, the set of ILR, hereinafter denoted as $\boldsymbol{r} = \{r_1, r_2, \cdots, r_{n_e}\}$ where n_e is the total number of exposures, uniquely determines the respective *infection link* (i.e., source of infection) for every exposure. The distribution of $\boldsymbol{r}$ can be shown to be $U(0,1)$ under their construction scheme and the model assumption given by Equation 4.1 and is independent *a priori* of the form of the spatial kernel. Its posterior samples, hereinafter denoted as $\tilde{\boldsymbol{r}}$, can be easily imputed in standard data augmentation algorithms such as Markov chain Monte Carlo (MCMC) by inverting the construction procedures of ILR and imputing the infection links. On applying a classical test to $\tilde{\boldsymbol{r}}$ for its compliance with $U(0,1)$, a posterior distribution of *p-values* is generated from which the evidence against the model assumption can be discerned. Specifically, we measure the evidence against the model by $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|\boldsymbol{y})$, the proportion of the posterior p-values which are less than 0.05. *Anderson-Darling* hypothesis test (Lewis, 1961) is adopted (for details see Chapter 3). We consider the six-cluster epidemic data mentioned in the last section.

In previous sections we have shown that increased availability of genetic data improves the estimation of the transmission graph. Given that the imputations of $\tilde{\boldsymbol{r}}$ rely on the imputed infection links (equivalently the transmission graph), increased availability of genetic data may potentially increase the sensitivity of the test based on $\tilde{\boldsymbol{r}}$ over the mis-specification of the model. Table 4.10 shows that this improvement of sensitivity is indeed achieved. In Table 4.10 we notice that when an incorrect spatial kernel

Table 4.10: Values of $\pi(P(\widetilde{\boldsymbol{r}}) < 0.05|\boldsymbol{y})$, where $\boldsymbol{y}$ is the observed data, estimated from 2,000 posterior samples of ILR computed from the six-cluster epidemic under different model assumptions regarding the spatial kernel (noted that Kernel A is the correct spatial kernel)

	$\pi(P(\widetilde{\boldsymbol{r}}) < 0.05\|\boldsymbol{y})$			
Sampling	100%	80%	50%	0%
Kernel A	10%	10%	9%	13%
Kernel B	50%	39%	36%	28%
Kernel C	100%	100%	100%	78%

(Kernel B or Kernel C) is used, stronger evidence against the model is discerned as more genetic data become available. Note that a power-law kernel (Kernel B) and the exponentially-bounded kernel (Kernel A) appear to be more capable at mimicking each other, particularly when there are not many observations corresponding to very-short and very-long transmissions, and hence the test sensitivity is lower in general compared to the case of fitting Kernel C. Also note that when the correct kernel is used, no significant evidence against the model is elicited.

4.6 Case study: spread of foot-and-mouth disease virus in UK (Darlington, Durham county, 2001)

In this section we apply our algorithm to a localised FMDV outbreak that occurred in the UK (Darlington, Durham County) in 2001 in which 12 infected premises were observed and each premises was sampled to obtain one virus sequence (Cottam et al, 2008; Morelli et al, 2012) with sequence length of sequence $n = 8176$. The geographical locations, the sampling times and removal (i.e., culling) times of the infected premises were reported. Estimated onset dates of lesions were also provided by experts at the times of sampling. The data were previously analysed by Morelli et al (2012) in one of the first important attempts to jointly consider epidemiological and genetic data in a dynamic framework. They adopted a pseudo-likelihood approach which does not account for unobserved transmitted sequences. Here we analyse the data using our methodology.

Similar to previous sections, we fit a spatial SEIR model to data. In particular, we assume that sojourn times in classes E and I follow $Gamma(a, b)$ characterised by the shape a and scale b and $Exp(\mu)$ characterised by the mean μ respectively; and the spatial kernel is assumed to be $\exp(-\kappa d_{ij})$. The model is fitted to the data using methods as described in Section 4.2.5. We consider whole genome sequencing in this

section. The estimated onset dates of lesions provide important information on the starting dates of infectiousness for infected premises as these two dates were suggested to be close to each other (Charleston et al, 2011). To incorporate uncertainty in the estimated lesion onset dates, for each infected premises we allow the infectious times to vary within a 14-day interval centered at the estimated lesion onset dates provided. It is noted that, given that the maximum of the estimated durations between lesion onset times and sampling times is 7 days, 14 days may represent a conservative upper bound of the estimation uncertainty.

4.6.1 Revisiting the inference of the transmission dynamics

Inference for the mutation rates and the transmission graph

The mutation rates have a very significant contribution to the likelihood and therefore to the joint inference of epidemic and evolutionary process. Application of our method enables us to estimate the mutation rates (Figure 4.25 and Figure 4.26), which were assumed to be known in Morelli et al (2012). Note that we allow two types of mutation (transition and transversion) while previous analyses (Cottam et al, 2008; Morelli et al, 2012) assumed a single aggregate mutation rate. Nevertheless, the orders of magnitude of our estimated mutation rates are consistent with the literature (Cottam et al, 2008; Mettenleiter and Sobrino, 2008).

Figure 4.27 shows the transmission graphs with the highest and second highest posterior probability. We first notice that our results validate the single-cluster assumption made by Morelli et al (2012). Similar to their analysis, the premises K was also identified as the index case of the transmission with high posterior probability. The longest transmission sub-path (i.e., $K \rightarrow F \rightarrow G \rightarrow I \rightarrow J$) coincides with their estimate. The most probable infection sources for premises O and L were also identical to ours. The infection sources of remaining premises were not entirely consistent with our estimate - for example, the sources for premises C and P were only identical to ours in our second most probable transmission graph, and the source for premises M was identified to be premises O instead of premises D. Nevertheless, the posterior of the transmission graph which we obtain is largely consistently with the earlier analysis, reinforcing the argument that pseudo-likelihood approach may be effective if the transmission network is of primary interest.

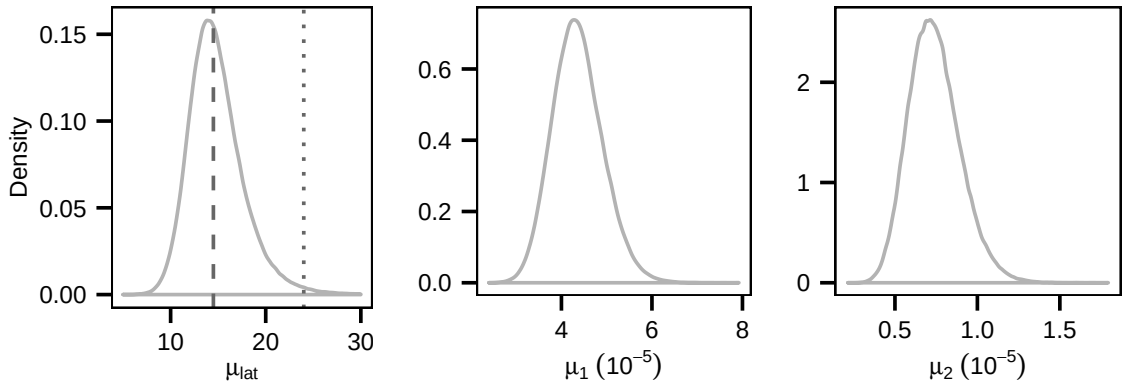


Figure 4.25: Posterior distributions of the mean latent period, denoted as μ_{lat}, and of the transition rate μ_1 and transversion rate μ_2. The grey dashed line and the dotted line indicate the median value of μ_{lat} obtained from our analysis and from Morelli et al (2012) respectively.

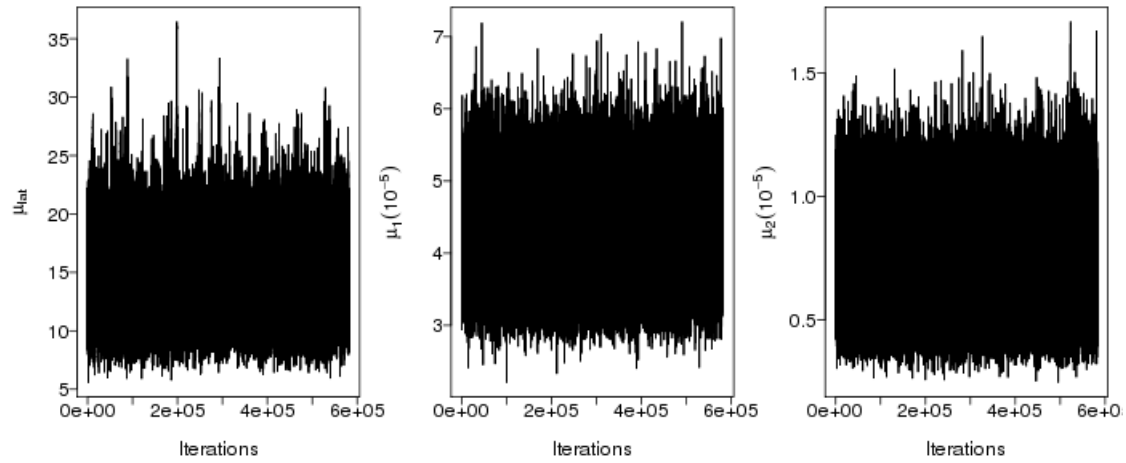


Figure 4.26: Traceplots for the posterior samples of the mean latent period and of the transition rate and transversion rate.

Inference for the latent period

The typical value of the latent period (i.e., sojourn times in class E) of FMD suggested in the literature is around 5 days (with 95% confidence interval [1,12]) (Charleston et al, 2011; Keeling et al, 2001). However, with the same dataset, the median of the mean latent period was estimated to be much higher (24 days with 95% C.I. [17 days, 35 days]) (Morelli et al, 2012). These authors hypothesised that the over-estimation was likely due to the scenario that some of the infected premises in the data were actually infected by undetected infectious premises. Figure 4.25 shows the posterior distribution of the mean latent period obtained using our method. It suggests a significantly lower median value of the mean latent period, 14.2 days, compared with the previous estimate of 24 days. Although our estimated mean latent period is much closer to the range suggested in the literature it is nevertheless distinctly high,

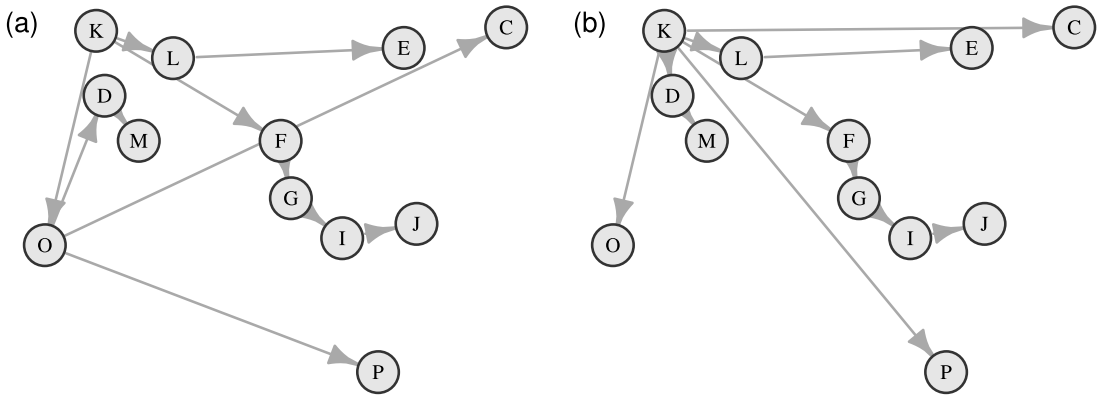


Figure 4.27: (a) The transmission graph with highest posterior probability, 0.89; (b) The transmission graph with the second highest posterior probability, 0.08. The same set of labels of premises used in Morelli et al (2012) are adopted for to facilitate comparison.

supporting the notion that undetected infected premises may play a role (Morelli et al, 2012).

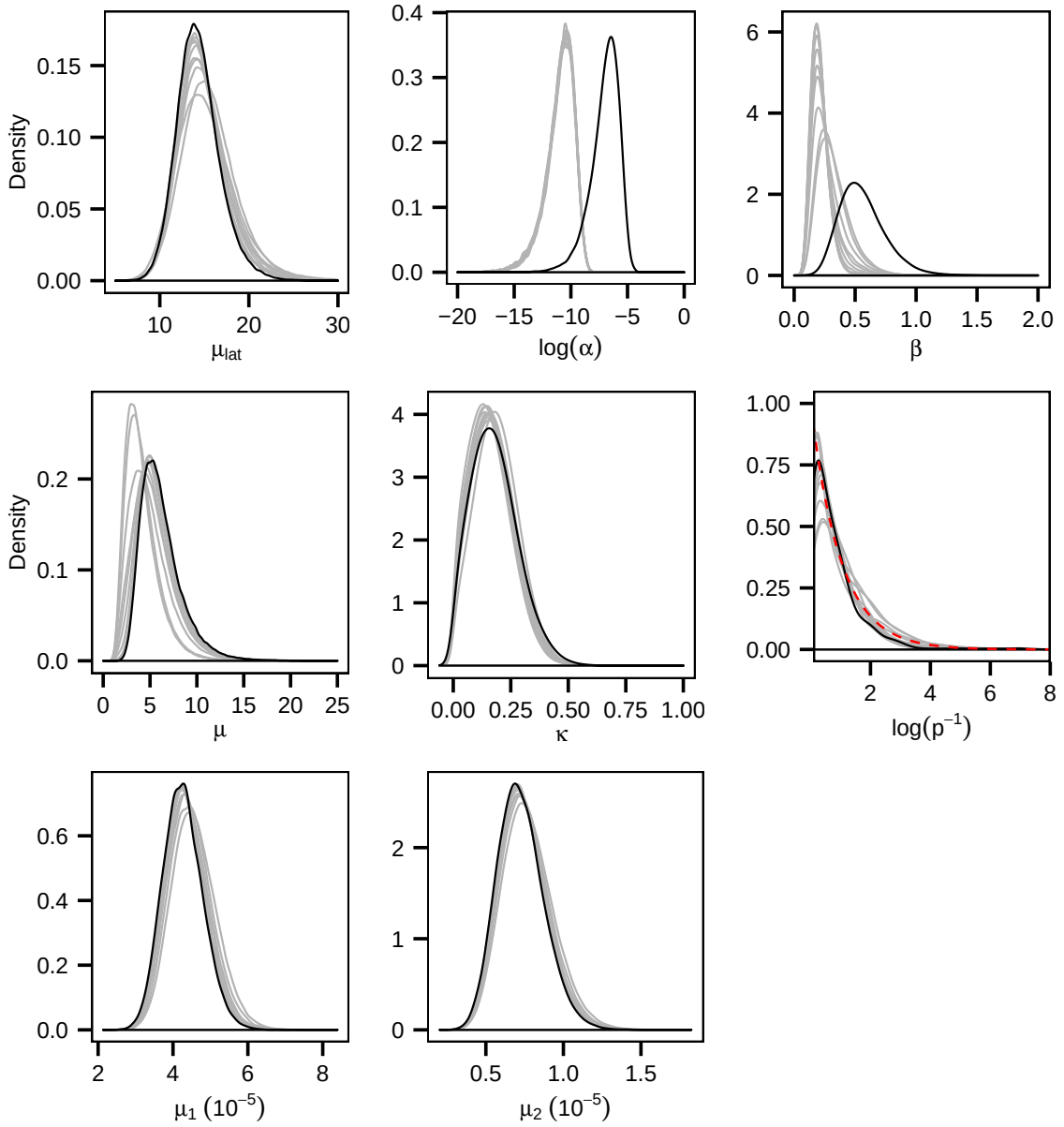


Figure 4.28: Posterior distributions of the full set of model parameters obtained from fitting the model to 10 independently simulated datasets obtained by adding 300 randomly assigned susceptible premises to the 2001 FMD data (grey curves). The posteriors corresponding to the case without considering susceptibles are coloured in black. Non-informative flat priors are used for model parameters. Note that the posterior distributions of p appear to be almost the same as its prior (i.e., $U(0,1)$) - to facilitate comparison, the posteriors of $log(p^{-1})$ are presented, and they look almost identical to an $Exp(1) = log(U(0,1)^{-1})$ represented by the red dotted line, which suggests that the data are not sufficient for estimating p (see more discussion in Section 4.7).

4.6.2 Inclusion of unreported susceptibles

The number and locations of susceptible premises in the region were not reported and therefore were not considered in the earlier analysis (Morelli et al, 2012). In this

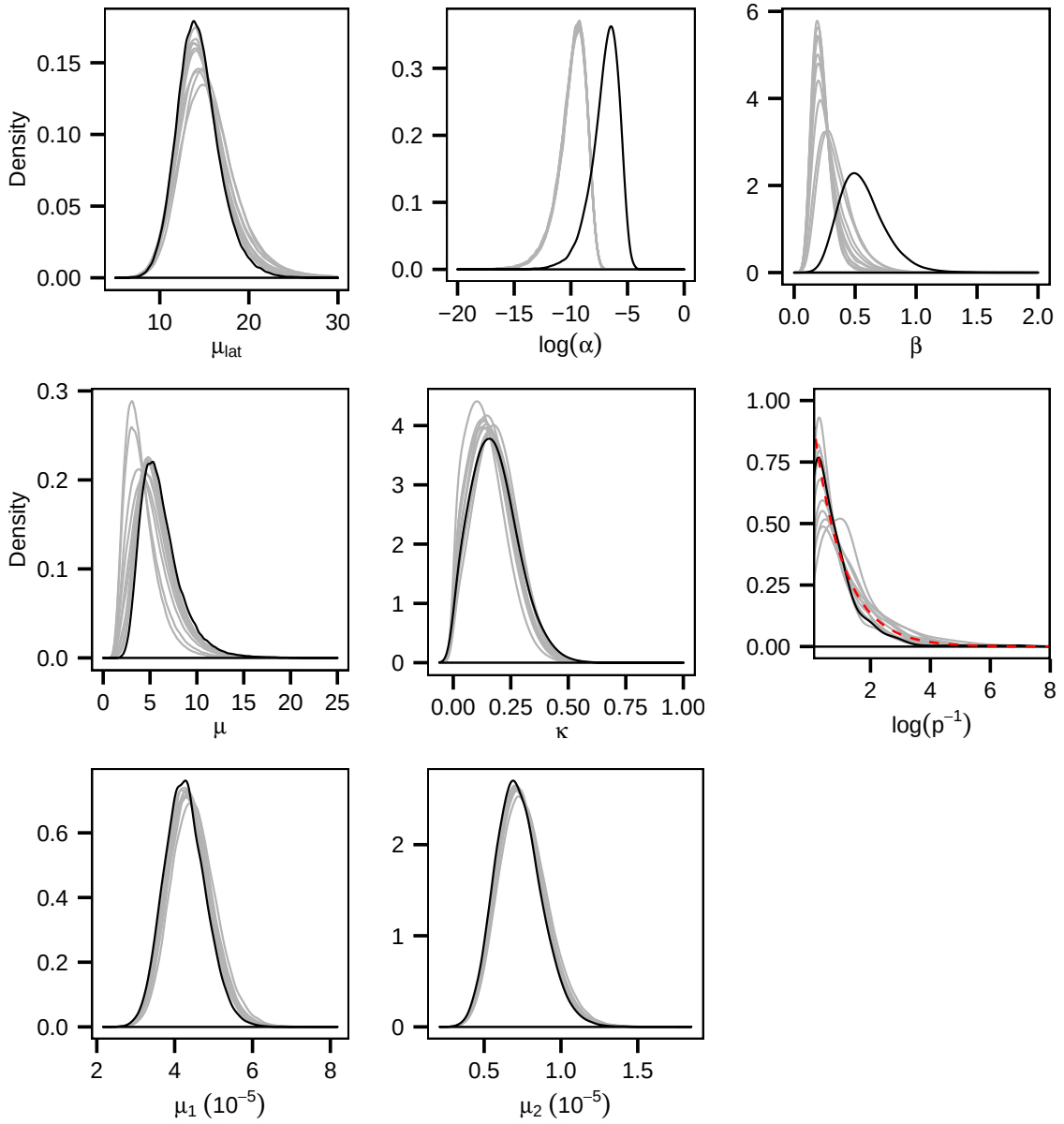


Figure 4.29: Posterior distributions of the full set of model parameters obtained from fitting the model to 10 independently simulated datasets obtained by adding 100 randomly assigned susceptible premises to the 2001 FMD data (grey curves).

section we investigate the effect of unreported susceptibles on estimation by first randomly assigning 300 susceptible premises in a minimal rectangular region (253 km^2) which encompasses the sampled premises. Note that the number of susceptible farms (300) we choose ensures that the farm density in the area we consider is consistent with the crude farm density across Durham County (Defra, 2009; Robinson, 2007). Results show that most of the model parameters, except the primary and secondary transmission rates, are robust to the inclusion of a considerable number of susceptible (Figure 4.28). In particular, we notice that the mean latent period is only slightly affected. The posterior distribution of the transmission graph is largely unaffected (not

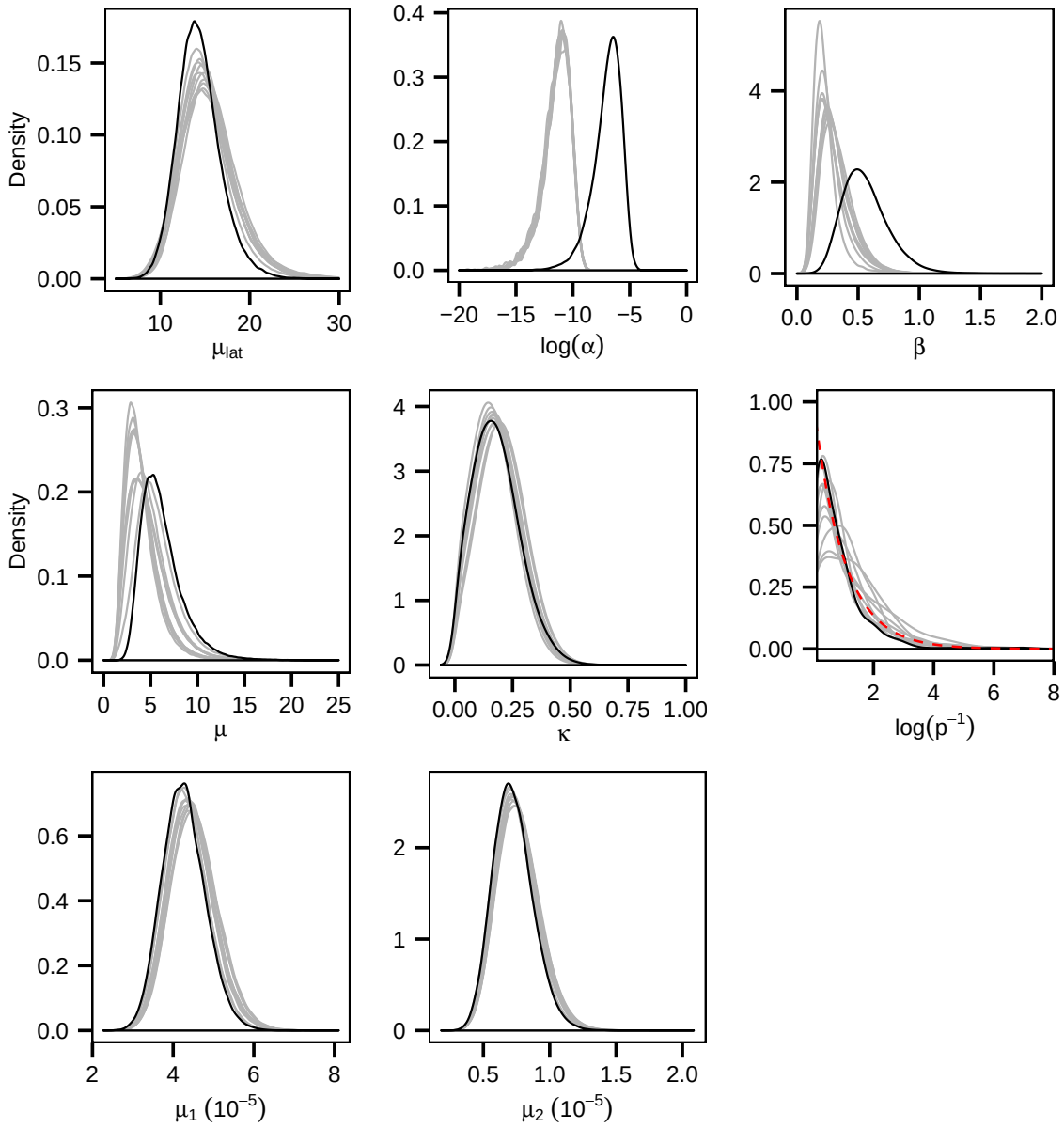


Figure 4.30: Posterior distributions of the full set of model parameters obtained from fitting the model to 10 independently simulated datasets obtained by adding 500 randomly assigned susceptible premises to the 2001 FMD data (grey curves).

shown). A lower level of farm density (100 susceptible premises) and a higher farm density (500 susceptible premises) are also considered. Figure 4.29 and Figure 4.30 suggest similar findings observed in Figure 4.28.

4.7 Validation of the methodology

4.7.1 Fitting a full model to epidemic data

In this section we perform a verification on the implementations of the part of our algorithm for sampling the unobserved sequences and transmission graphs (fitting the six-cluster epidemic simulated above). While estimation of the full set of model parameters is not feasible given insufficient genetic data, we consider a minimally sufficient case in which we compare the posterior distributions of the coverage rate and κ obtained respectively from two scenarios - in scenario I, we assume no genetic data and fit the full model (epidemic and genetic model); in scenario II, assume no genetic data and fit only the epidemic model. Assuming other model parameters to be known, we also impute the times of exposures in both scenarios, and in Scenario I we impute unobserved transmitted sequences, the master sequence and the transmission graphs which are the key components of our algorithm. Theoretically, the two scenarios should give identical posterior distributions as the observed data are the same. Figure 4.31 shows that the posterior distributions of the coverage rate and κ appear to have no significant difference, which in turn supports our algorithm. Note that the insignificant difference between the posterior (cumulative) distributions of the coverage rate is likely to be caused by numerical rounding behaviours due to the vast difference in model dimensions and hence in the magnitudes of likelihood values. In fact, the posterior (non-cumulative) densities suggest very similar coverages rates- in scenario I, the coverage rate has mean 0.68 and a standard deviation 0.027; in scenario II, the coverage rate has mean 0.68 and standard deviation 0.028.

4.7.2 Posterior distribution of parameter p for the FMD outbreak (Darlington, 2001)

Figure 4.28 shows that the posterior distributions of p look almost the same as its prior. Heuristically speaking, the data are informative about p when there are multiple clusters (i.e., when there are multiple background sequences S_1,S_2,... derived from the master sequence G_M). Since single-cluster transmission graphs are highly supported by our analysis, should our algorithm be efficient in exploring the sequence and tree space and correctly implemented we would expect the posterior p to be identical to its prior as the data give no extra information on p. Here we also state a mathematical argument to support our notion above.

Proposition 4.7.1 *Conditioning on a single-cluster transmission graph and on each*

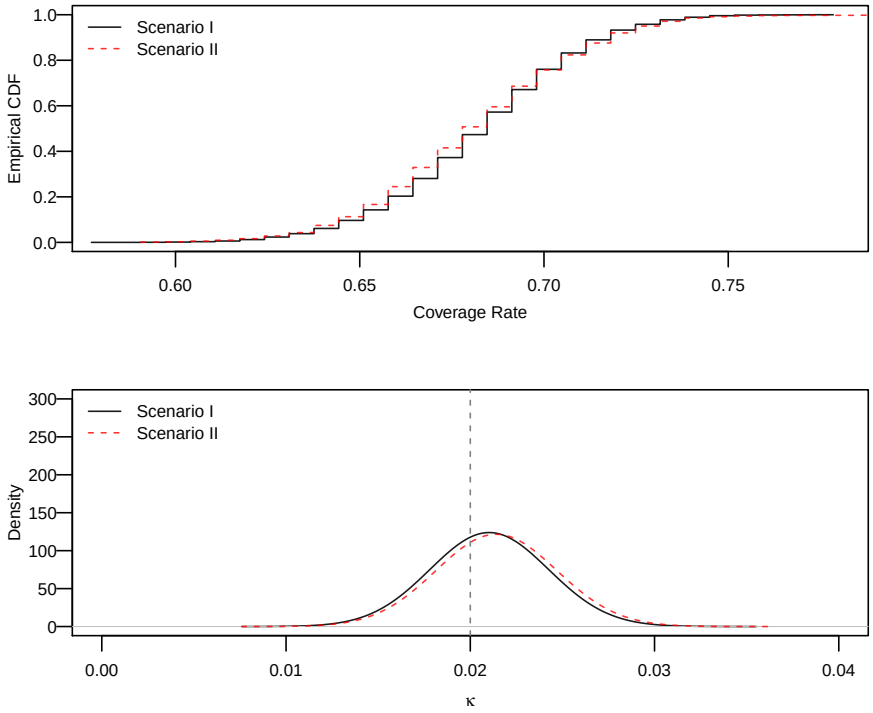


Figure 4.31: Comparisons between the posterior distributions of the coverage rates and of κ obtained from fitting two models, the full model (Scenario I) and the epidemic model (Scenario II), to the epidemic data (no sampled sequences).

base of the master sequence G_M is being proposed uniformly from the set $\omega_N = \{A, C, G, T\}$, the posterior of p would be identical to its prior.

Proof Denoting $\pi(p)$ and $\pi(p|S)$ as the the prior and the posterior distribution (given the only one background sequence S for the cluster) of p respectively, we have

$$\begin{aligned}
\pi(p|S) &\propto \pi(p) \times \sum_{G_M} P(G_M) \times P(S|p, G_M) \\
&\propto \pi(p) \times \sum_{G_M} P(S|p, G_M) \quad (\because P(G_M) = constant) \\
&= \pi(p) \times \sum_{G_M} P(G_M|p, S) \\
&= \pi(p)
\end{aligned}$$

The last equality holds as

$$\sum_{G_M} P(G_M|p, S) = 1. \quad \blacksquare$$

Note that when there are more than one cluster, the second equality does not hold (i.e., $P(S_1, S_2, \dots | p, G_M) \neq P(G_M | p, S_1, S_2, \dots)$).

4.8 Discussion

In response to the increasing availability of genetic data of pathogens in epidemic outbreaks, substantial progress has also been made in the realm of the joint analysis of epidemiological and genetic data (Ypma et al, 2012, 2013; Morelli et al, 2012; Jombart et al, 2014; Cottam et al, 2008). The literature has focused, successfully, on imputing the transmission tree among the population by adopting pseudo-likelihood like approaches. These approaches, however, do not take the unobserved transmitted sequences into account and make some key simplifying assumptions on the transmission dynamics, such as assuming independence among subtrees of transmission, and moreover they focused only on the single-cluster epidemic. Unequivocally, other model parameters such as the timing of exposures, the transmission rate and the spatial kernel parameters also play a key role in predicting, managing and controlling the epidemics (Parry et al, 2014; Lau et al, 2014b; Ster et al, 2009; Ferguson et al, 2001) but have received less attention. Therefore, there is a need for methods that can more comprehensively capture both the epidemiological and evolutionary process. However, a more exact analysis of the joint process requires consideration of the unobserved transmitted sequences, which in general is hindered by the vast model dimension mainly contributed by the combination of the transmission graph space with the sequence space.

Primarily, we show that an efficient joint imputation of the graph and transmitted sequences is feasible, and the epidemiological and evolutionary model parameters can be reasonably recovered. A key feature arising from imputing the transmitted sequence is that exposures without a sampled sequence can be naturally incorporated into the transmission dynamics in contrast to pseudo-likelihood approaches which only consider sampled exposures. Our method is also validated by computer experiments and mathematical arguments. Results show that the transmission graph is more accurately estimated when more genetic data are available. Particularly, in the scenario with full sampling the transmission graph can be fully or almost fully recovered. Genetic data may also greatly improve the precision of the estimation of the epidemiological model parameters, the secondary transmission rate and the spatial kernel parameter in particular.

The more general scenario with an unknown number of clusters of transmission has not been well addressed with respect to joint analysis of epidemiological and genetic

data. We explore the scenario that background sequence is a random variant of a so-called (unobserved) master sequence and we pursue the imputation of this master sequence and the parameter which governs the random variation process. Results show that the imputations of these unknown quantities are feasible and they can be accurately estimated.

Results also support the use of partial genome data, bearing important implications for future study designs. For example, we show that it can be desirable, for both economic and computational purposes, to consider partial genome sequencing if the primary interests are the transmission graph and the epidemiological parameters. For pathogens with very high mutation rates, the estimation of full set of model parameters may be feasible given only a relatively small percentage of sub-sampling of the exposed cases if the latent period distribution is assumed to be known. With moderate to high mutation rates, a small percentage of sub-sampling of exposed cases may be sufficient for full estimation. Also, clusters may be identified when only a small sub-sampling percentage of cases are sampled.

The form of the spatial kernel has an important implication on the prediction and control of epidemic outbreaks and in Chapter 3 we have shown that mis-specification the spatial kernel can be detected through the use of imputed residuals. In this chapter we further show that increased availability of genetic data may reinforce the sensitivity of this test on the mis-specification of the spatial kernel, particularly when the observed epidemic data alone do not suggest much difference between competing models.

Application of our method to a FMD spread in 2001 in the UK demonstrates that a more realistic estimate of the latent period is achieved, compared to those in earlier literature, and that the mutation rates can be estimated. This highlights the importance of explicitly taking into account the transmitted sequences for constructing a more accurate and integrated representation of the transmission dynamics, aiming at reliable prediction and effective management of disease outbreaks.

There are limitations to our work. For instance, by assuming a dominant strain on an exposure at any time point, we have not considered the within-host dynamic of the pathogens which renders our framework more appropriate for acute infectious pathogens or for situations with a narrow transmission bottleneck. Nevertheless, as pointed out in Ypma et al (2013), the within-host effective pathogen population size and pathogen generation time required to estimate the within-host dynamic may not be available in general. A neutral evolution model without accounting for any selection pressure is assumed. Also, we have not considered heterogeneity over sites which may be represented by introducing additional mutation parameters. Related

potential future work will be discussed in Chapter 5. We have, nevertheless, shown that our algorithm works under a various realistic scenarios and can enhance our understanding to the transmission dynamics of a real-world epidemic.

Chapter 5

Conclusion and future work

5.1 Conclusion to the thesis

The work in this thesis represents advances in addressing two key challenges in epidemiological and ecological modelling: the lack of an effective and easily deployable model-assessment tool and a statistically sound joint inferential framework for epidemic and evolutionary processes.

As discussed in previous chapters, predicted dynamics of epidemiological and ecological systems can be extremely sensitive to the choice of model, with consequent implications for the design of control strategies (Ster et al, 2009; Ferguson et al, 2001; Fraser et al, 2004; Filipe and Maule, 2004). In particular, it is well-known that the spatial transmission mechanisms are difficult to assess in practice yet have major implications for optimal control strategies. Earlier work (Box, 1980) has championed the view that Bayesian and classical reasoning are natural approaches to follow for parameter estimation and model criticism respectively and therefore should be used in combination. In Chapter 3 we have presented a statistical framework that combines classical and Bayesian reasoning in testing for mis-specifications of a spatio-temporal model by investigating the consistency of imputed latent residuals with a known sampling distribution using a classical hypothesis test. The latent-residual testing framework proposed enables targeted assessments to specified components of general spatio-temporal process (including the spatial kernel) as each type of residual is designed to be sensitive to a particular model component. The idea of reconstructing such processes with latent residuals is indeed very general and modellers are not restricted to the residuals processes proposed. That is they may design other residual processes that serve their own purposes. For example, while in Chapter 3 we have designed the Infection-link Residuals to distinguish kernels with distinct properties

at short and long distances, alternative designs of the residuals may be considered if the goal is to distinguish other properties (see discussion in 3.4.1). We have also shown that our approach can inform modellers of the nature of the model misspecification and not just the degree of mis-specification, thus enabling a finer diagnosis than that is possible with conventional model assessment tools. Analysis of simulated epidemics and data on the spread of an invasive species in the UK demonstrates how our model testing framework can be utilised to diagnose the model fit and how the results can be interpreted in practice. The fine model diagnostics enabled by our framework could lead to a better risk assessment of future spread and therefore more appropriately targeted control measures, which results in more efficient management of disease by reducing costs and harm they cause. A R package *EpiResTest* (Lau and Pollock, 2014) is built (beta version) to implement the latent-residual tests developed in Chapter 3 (see details in Appendix A).

Substantial progress has been made recently on integrating epidemic and genetic data to infer the transmission network in an epidemic. The use of pseudo-likelihood has proved effective in estimating the transmission network. However, there has been less progress on methods of robust inference for other aspects of the transmission dynamics which are important in predicting and managing disease outbreaks. In Chapter 4 we show that unobserved transmitted sequences can be imputed effectively and the transmission dynamics can be reasonably recovered in the presence of exposures without observed sequences. We also show that increased availability of genetic data can aid the estimation of epidemiological model parameters, and we demonstrate the value of partial genomic data in quantifying outbreaks which has important implications for sampling designs of future studies. Moreover, we demonstrate that genetic data may enhance our ability to detect mis-specification of the spatial transmission mechanism when they are used in combination with the residual methods of Chapter 3. The proposed framework is subsequently applied to analyse a localised spread of foot-and-mouth disease virus in the UK and we show that understanding to the transmission dynamics can be greatly enhanced.

Potential future developments are discussed now.

5.2 Future developments for the residual testing

5.2.1 A sequential approach

In Chapter 3 we have proposed a sequential latent-residual testing approach which may be more sensitive than the non-sequential approach presented earlier in that

chapter, when there is more evidence of model mis-specification contained in the early observations of an epidemic. However, from the point of view of a classical observer, it is still not appropriate to claim that this (sequential) testing procedure has a known type I error, for the critical values c_i derived for the sequential approach are required to be re-computed when the sample size of the epidemic changes (i.e., to a classical observer, the testing result still depends on the random realisation of the number of exposures in the epidemic). One possible solution may be to consider the classic sequential likelihood ratio test proposed by Wald (1973) where we are not required to pre-specify the testing levels of the sample size. However, to use Wald's approach, we have to specify a distribution model of the residual process in the alternative hypothesis, which may render it potentially suitable only when there is a certain type of model mis-specification in mind to be tested (so that the alternative may be formulated). Also, in this sequential approach the choice of the number and size of the m subsets of the full residual sample can be subjective and arbitrary, there may be an optimal subsetting scheme (e.g., by relating it to the instantaneous growth rates of an epidemic), which requires further investigations.

5.2.2 Exposure Time Residuals (ETR)

Scrutiny of ETR has not offered much potential to detect mis-specification of the model in the scenarios considered in Chapter 3. The confounding effects (i.e., the test sensitivity relies on the assumption of other model components not being assessed) we explored in 3.7.1 may actually point to a further research direction − as the posterior samples of ETR summarise the entire model structure (see the construction of ETR in section 3.3.2), it should be more prone to confounding effects. ETR may become sensitive and useful under scenarios where the sources of such confounding effects are limited − for example, future work may investigate how it performs when the latent period distribution, which is believed to have a direct effect on the exposure time, is assumed to be known. Also, ETR might become useful when the assumption of $Exp(1)$ threshold no longer holds (Streftaris and Gibson, 2012).

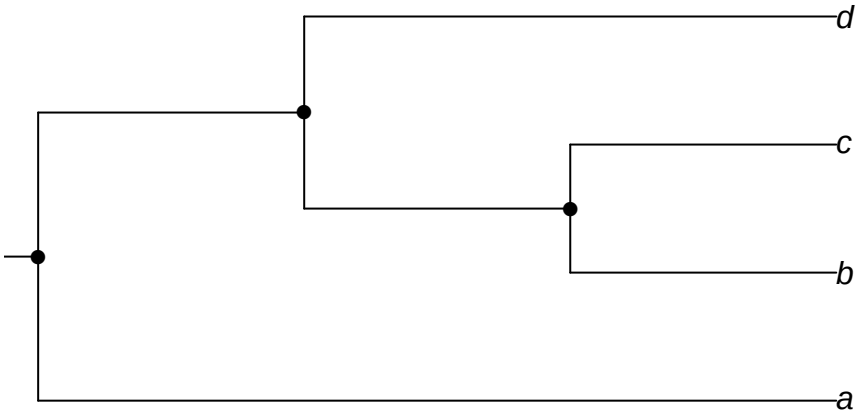
5.3 Future developments for the joint analysis of epidemic and genetic data

5.3.1 Modelling within-host dynamics

As discussed in 2.2.4, the evolutionary model considered in Chapter 4 is an abstraction of a much more complex biological world. A key extension is to model the within-host diversity of the pathogens which has not been captured by Equation 4.2.

We can consider a *Yule process* (Yule, 1925) (a binary branching process) which asserts that each strain has a constant rate λ at which it gives birth to a new strain. Figure 5.1 illustrates a realisation of the Yule process starting with a single strain as it evolves to four strains at time t.

Figure 5.1: An illustration of a Yule process. Starting from one strain, after going through 3 birth events at 3 nodes (indicated by the black dots), evolves to 4 strains ($a,\ b,\ c,\ d$) at time t.



Probabilistic distribution of the number of strains n_t

Let $\mathbf{t_d} = (t_{d_1}, t_{d_2}, \ldots, t_{d_{n_t-1}})$ be the vector of time between the respective nodes (i.e., the divergent times) and t, where n_t is the total number of strains at time t (e.g.,

n_t=4 in Figure 5.1).

It can be shown that n_t (Cox and Lewis, 1966; Rannala, 1997) follows a negative binomial distribution with parameters n_0 and $1 - e^{-\lambda t}$. That is,

$$n_t|n_0, \lambda, t \sim NB(n_0, 1 - e^{-\lambda t}) = \binom{n_t - 1}{n_t - n_0} e^{-n_0 \lambda t} (1 - e^{-\lambda t})^{n_t - n_0}. \tag{5.1}$$

where n_0 is the initial number of strains at time zero.

Probabilistic distribution of divergent time t_{d_i}

Another distribution of interest is $t_{d_i}|n_0, \lambda, t$, the distribution of the time between the i^{th} divergent time and t. It can be first shown (Nee, 2001) that

$$P(\mathbf{t_d}|n_t, n_0, \lambda, t) = (n_t - n_0)!(\frac{\lambda}{1 - e^{-\lambda t}})^{n_t - n_0} e^{-\lambda \sum_{n_0+1}^{n_t} t_{d_i}}. \tag{5.2}$$

Equation 5.2 above corresponds to the probability density function of the order statistics of $n_t - n_0$ independent and identically distributed random variables with *truncated exponential distributions* which are independent of n_t and n_0. As a result (Nee, 2001), we see that Equation 5.2 implies

$$P(t_{d_i}|\lambda, t) = \frac{\lambda e^{-\lambda t_{d_i}}}{1 - e^{-\lambda t}}. \tag{5.3}$$

Suppose we assume each strain on an infected premises at time t has an equal probability to be transmitted to a susceptible premises and assume the initial infection with a single strain, the subsequent growth of strains within each premises can be modelled by Equation 5.1. A foolproof but maybe computationally unmanageable approach is to jointly impute the transmission tree and the branching patterns in Figure 5.1 within each infective and infected premises. Consider two sequences sampled from two premises. In fact, it may be sufficient to just know (impute) their *most recent common ancestor (MRCA)*, which corresponds to either a node or a tip in Figure 5.1, to describe adequately their evolutionary relationship. Although Equation 5.2 and Equation 5.3 allow us to work out the times of the nodes (hence the times for MRCA, t_{MRCA}), further theoretical developments are needed. For example, given any pairs of descended samples (e.g., a and d in Figure 5.1), we need to know the probability they have a particular MRCA so we can assign the respective t_{MRCA}. However, as far

as we are aware, this probability depends on the branching patterns and presents a complicated tree-combination problem to which a solution is not available so far (e.g., Mulder (2011); Steel and McKenzie (2001)).

Existing approaches cannot be used directly for the sake of an accurate joint inferential framework. For example, Ypma et al (2013) used a simple pathogen-effective-size growth model but assumed it is completely known. Didelot et al (2014) constructed the phylogeny in hosts independently of the transmission network opposed to a truly joint approach. The feasibility of extending these approaches, for example, estimating the growth model used by Ypma et al (2013) jointly with the transmission dynamics, requires further research. In Chapter 4 we have used a universal master sequence G_M coupled with a variation process to model the background infection process and shown that the master sequence and the variation process can be accurately estimated together with the transmission dynamics. It may be then possible that the within-host diversity can be modelled in a similar manner, where G_M would now represent the "local" master sequence within the host.

5.3.2 Alternative sampling schemes of genetic data

Lastly, we have considered random sub-sampling of exposures for sequence data in Chapter 4. However, our framework is not restricted to that and alternative sampling schemes may be considered and investigated. For instance, in events of superspreading where many infections may occur (cluster) during a short period of time, we may accordingly consider a more concentrated sampling (e.g., at the peak of incidence) rather than an "even" (random) sampling as most of the information of the evolutionary process may be contained in the observations during this period of superspreading.

Appendix A

A R package for latent residuals test

A.1 A brief description

A R package *EpiResTest* (Lau and Pollock, 2014) is built (beta version) to implement the latent-residual tests developed in Chapter 3 which are specifically designed to measure the goodness-of-fit of different model components of a general spatial SEIR model commonly used in epidemiology and ecology studies. Functions in the package require inputs of snapshots of posterior samples of model parameters (e.g., exposure times) and impute the residuals. They do not compute any summary statistics such as the posterior p-value used in Chapter 3 so users may analyse the raw distributions of the residuals. The underlying functions are coded in C++ so they should be generally quick.

A.2 Flexibility of the package

SEIR models have been considered in this thesis but more restrictive model classes such as SIR and SI can be accommodated by this package. Note that although each test is specifically designed to be sensitive to one particular component of a spatio-temporal model, they are not theoretically restricted to only testing these specific aspects and can be used as a general model assessment tool like any other conventional model selection techniques. Also, as functions in the package only compute the raw residuals, they can be readily used for any potential sequential approaches such that that we have proposed in 3.7.2.

In Chapter 3 the transmission network is not explicitly modelled and infection links are only imputed for the sake for computing the Infection-link Residuals, unlike in Chapter 4 where the transmission network is explicitly modelled and imputed as model parameters. As the transmission network is not always a necessity in epidemiological modelling, we provide two versions of the function for imputing the Infection-link Residuals.

Bibliography

Aitkin M (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. Statistics and Computing 7(4):253–261

Aitkin M, Boys RJ, Chadwick T (2005) Bayesian point null hypothesis testing via the posterior likelihood ratio. Statistics and Computing 15(3):217–230

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88(422):669–679

Anderson RM (1988) The epidemiology of HIV infection: variable incubation plus infectious periods and heterogeneity in sexual activity. Journal of the Royal Statistical Society Series A 151:66–93

Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society Series B 72(3):269–342

Bailey NT, et al (1975) The mathematical theory of infectious diseases and its applications. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 41:379–406

Bellosta C (2011) Adgoftest: Anderson-Darling GoF test. R package version 03, URL http://CRAN R-project org/package= ADGofTest

Berger JO, Bernardo JM (1992) On the development of reference priors. Bayesian Statistics 4(4):35–60

Bernardo JM (1979) Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society Series B pp 113–147

Bernardo JM (2005) Reference Analysis. Handbook of Statistics 25(188):17–90

Bhowmik PC (2005) Characteristics, significance, and human dimension of global

invasive weeds. In: Invasive plants: ecological and agricultural aspects, Springer, pp 251–268

Bivand R, Anselin L, Berke O, Bernat A, Carvalho M, Chun Y, Dormann C, Dray S, Halbersma R, Lewin-Koh N, et al (2011) spdep: Spatial dependence: weighting schemes, statistics and models

Bolker B, Grenfell B (1995) Space, persistence and dynamics of measles epidemics. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 348(1325):309–320

Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society Series A (General) pp 383–430

Casella G, George EI (1992) Explaining the Gibbs sampler. The American Statistician 46(3):167–174

Catterall S, Cook AR, Marion G, Butler A, Hulme PE (2012) Accounting for uncertainty in colonisation times: a novel approach to modelling the spatio-temporal dynamics of alien invasions using distribution data. Ecography 35(10):901–911

Cauchemez S, Ferguson NM (2008) Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london. Journal of the Royal Society Interface 5(25):885–897

Celeux G, Forbes F, Robert CP, Titterington DM, et al (2006) Deviance information criteria for missing data models. Bayesian Analysis 1(4):651–673

Charleston B, Bankowski BM, Gubbins S, Chase-Topping ME, Schley D, Howey R, Barnett PV, Gibson D, Juleff ND, Woolhouse ME (2011) Relationship between clinical signs and transmission of an infectious disease and the implications for control. Science 332(6030):726–729

Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. The American Statistician 49(4):327–335

Cook A, Marion G, Butler A, Gibson G (2007a) Bayesian inference for the spatio-temporal invasion of alien species. Bulletin of mathematical biology 69(6):2005–2025

Cook AR, Otten W, Marion G, Gibson GJ, Gilligan CA (2007b) Estimation of multiple transmission rates for epidemics in heterogeneous populations. Proceedings of the National Academy of Sciences 104(51):20,392–20,397

Cook AR, Gibson GJ, Gilligan CA (2008) Optimal observation times in experimental epidemic processes. Biometrics 64(3):860–868

Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT (2008) Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proceedings of the Royal Society B: Biological Sciences 275(1637):887–895

Cox D, Lewis P (1966) The statistical analysis of series of events. John Wiley and Sons

Cox DR, Snell EJ (1968) A general definition of residuals. Journal of the Royal Statistical Society Series B (Methodological) pp 248–275

Dawe NK, White ER (1979) Giant cow parsnip (Heracleum mantegazzianum) on Vancouver Island, British Columbia. Canad Field-Nat 93(1):82–83

Dawid AP, Stone M (1982) The functional-model basis of fiducial inference. The Annals of Statistics pp 1054–1067

Debanne SM, Bielefeld RA, Cauthen GM, Daniel TM, Rowland DY (2000) Multivariate Markovian modeling of tuberculosis: forecast for the United States. Emerging Infectious Diseases 6(2):148

Defra (2009) Archive: Defra Economics and Statistics - June survey of agriculture and horticulture. `http://archive.defra.gov.uk/evidence/statistics/foodfarm/landuselivestock/junesurvey/results.htm`, accessed: 2014-07-02

Dempster AP (1974) The direct use of likelihood for significance testing. Proc Conf Foundational Questions in Statistical Inference pp 335–352

Dempster AP (1997) The direct use of likelihood for significance testing. Statistics and Computing 7(4):247–252

Didelot X, Gardy J, Colijn C (2014) Bayesian inference of infectious disease transmission from whole genome sequence data. Molecular Biology and Evolution 31(7):1869–1879

Diggle PJ, Gratton RJ (1984) Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society Series B (Methodological) pp 193–227

Ferguson NM, May RM, Anderson RM (1997) Measles: Persistence and synchronicity in disease dynamics. Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions pp 137–157

Ferguson NM, Donnelly CA, Anderson RM (2001) Transmission intensity and im-

pact of control policies on the foot and mouth epidemic in Great Britain. Nature 413(6855):542–548

Filipe JAN, Maule MM (2004) Effects of dispersal mechanisms on spatio-temporal development of epidemics. Journal of Theoretical Biology 226(2):125–141

Fraser C, Riley S, Anderson R, Ferguson N (2004) Factors that make an infectious disease outbreak controllable. Proceedings of the National Academy of Sciences 101(16):6146

Gelfand A, Ghosh S (1998) Model choice: A minimum posterior predictive loss approach. Biometrika 85(1):1–11

Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. Environment and Planning A 23(9):1269–1277

Geweke J, et al (1991) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, vol 196. Federal Reserve Bank of Minneapolis, Research Department

Gibson G, Otten W, N Filipe J, Cook A, Marion G, Gilligan C (2006) Bayesian estimation for percolation models of disease spread in plant populations. Statistics and Computing 16(4):391–402

Gibson GJ (1997) Investigating mechanisms of spatiotemporal epidemic spread using stochastic models. Phytopathology 87(2):139–146

Gibson GJ, Renshaw E (1998) Estimating parameters in stochastic compartmental models using Markov chain methods. Mathematical Medicine and Biology 15(1):19–40

Gilks WR (2005) Markov chain Monte Carlo. Wiley Online Library

Gomes MGM, White LJ, Medley GF (2004) Infection, reinfection, and vaccination under suboptimal immune protection: epidemiological perspectives. Journal of Theoretical Biology 228(4):539–549

Gordon NJ, Salmond DJ, Smith AF (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In: IEE Proceedings F (Radar and Signal Processing), IET, vol 140, pp 107–113

Gottwald TR (2010) Current epidemiological understanding of Citrus Huanglongbing. Annual Review of Phytopathology 48:119–139

Gottwald TR, Graham JH, Schubert TS (2002) Citrus canker: the pathogen and its impact. Plant Health Progress 10

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732

Grenfell B, Harwood J (1997) (meta) population dynamics of infectious diseases. Trends in Ecology Evolution 12(10):395–399

Grenfell B, Bjørnstad O, Kappey J (2001) Travelling waves and spatial hierarchies in measles epidemics. Nature 414(6865):716–723

Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303(5656):327–332

Hall CB, Douglas RG, Geiman JM, Meagher MP (1979) Viral shedding patterns of children with influenza b infection. Journal of Infectious Diseases 140(4):610–613

Han C, Carlin BP (2001) Markov chain Monte Carlo methods for computing Bayes factors. Journal of the American Statistical Association 96(455):1122–1132

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109

Jeffreys H (1935) Some tests of significance, treated by the theory of probability. In: Proceedings of the Cambridge Philosophical Society, Cambridge Univ Press, vol 31, pp 203–222

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences 186(1007):453–461

Jeffreys H (1961) Theory of probability. Oxford University Press, USA

Jégat C, Carrat F, Lajaunie C, Wackernagel H (2008) Early detection and assessment of epidemics by particle filtering. In: geoENV VI–Geostatistics for Environmental Applications, Springer, pp 23–35

Jombart T, Eggo R, Dodd P, Balloux F (2010) Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106(2):383–390

Jombart T, Didelot X, Cauchemez S, Viboud FC, Ferguson N (2014) Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Computational Biology DOI 10.1371/journal.pcbi.1003457

Kass RE, Raftery AE (1995) Bayes factors. Journal of the American statistical association pp 773–795

Kawaguchi I, Sasaki A, Boots M (2003) Why are dengue virus serotypes so distantly related? Enhancement and limiting serotype similarity between dengue virus strains. Proceedings of the Royal Society of London Series B: Biological Sciences 270(1530):2241–2247

Keeling M, Woolhouse M, May R, Davies G, Grenfell B (2002) Modelling vaccination strategies against foot-and-mouth disease. Nature 421(6919):136–142

Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT (2001) Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 294(5543):813–817

Keeling MJ, Woolhouse M, May RM, Davies G, Grenfell BT, et al (2003) Modelling vaccination strategies against foot-and-mouth disease. Nature 421(6919):136–142

Kimura M (1984) The neutral theory of molecular evolution. Cambridge University Press

Kleczkowski A, Gilligan C (2007) Parameter estimation and prediction for the course of a single epidemic outbreak of a plant disease. Journal of The Royal Society Interface 4(16):865–877

Kullback S (1987) The Kullback-Leibler distance. American Statistician 41(4):340–340

Lau MSY, Pollock J (2014) EpiResTest: Latent Residuals Test for Epidemiology and Ecology Models. URL `http://CRAN.R-project.org/package=EpiResTest`, r package version 1.0

Lau MSY, Marion G, Streftaris G, Gibson GJ (2014a) Bayesian inference in epidemics using new methodology for integrating epidemiological and genetic data, submitted for publication

Lau MSY, Marion G, Streftaris G, Gibson GJ (2014b) New model diagnostics for spatio-temporal systems in epidemiology and ecology. Journal of The Royal Society Interface 11:20131,093

Laud P, Ibrahim J (1995) Predictive model selection. Journal of the Royal Statistical Society Series B (Methodological) pp 247–262

Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proceedings of the National Academy of Sciences 96(19):10,752–10,757

Lewis PA (1961) Distribution of the Anderson-Darling statistic. The Annals of Mathematical Statistics pp 1118–1124

Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz W (2005) Superspreading and the effect of individual variation on disease emergence. Nature 438(7066):355–359

Marsaglia G, Marsaglia J (2004) Evaluating the Anderson-Darling distribution. Journal of Statistical Software 9(2):1–5

Mathews JD, McCaw CT, McVernon J, McBryde ES, McCaw JM (2007) A biological model for influenza transmission: pandemic planning implications of asymptomatic infection and immunity. PLoS One 2(11):e1220

Meng X (1994) Posterior predictive p-values. The Annals of Statistics pp 1142–1160

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21(6):1087–1092

Mettenleiter TC, Sobrino F (2008) Animal viruses: molecular biology. Horizon Scientific Press

Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Samule S (2012) A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Computational Biology 8:e1002,768

Morris M (1996) Behaviour change and non-homogeneous mixing. Models for Infectious Human Diseases: their Structure and Relation to Data, eds V Isham and G Medley pp 239–252

Mulder WH (2011) Probability distributions of ancestries and genealogical distances on stochastically generated rooted binary trees. Journal of Theoretical Biology 280(1):139–145

Muñoz A, Sabin CA, Phillips AN, et al (1997) The incubation period of AIDS. Aids 11(Suppl A):S69–S76

Neal P (2012) Efficient likelihood-free Bayesian computation for household epidemics. Statistics and Computing 22(6):1239–1256

Neal PJ, Roberts GO (2004) Statistical inference and model selection for the 1861 hagelloch measles epidemic. Biostatistics 5(2):249–261

Nee S (2001) Inferring speciation rates from phylogenies. Evolution 55(4):661–668

Neri FM, Cook AR, Gibson GJ, Gottwald TR, Gilligan CA (2014) Bayesian analysis

for inference of an emerging epidemic: citrus canker in urban landscapes. PLoS Computational Biology 10(4):e1003,587

O'Hagan A, Kendall MG, Forster J (2004) Kendall's Advanced Theory of Statistics: Bayesian Statistics. Vol. 2B. Arnold

O'Neill PD, Roberts GO (1999) Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 162(1):121–129

Ong JBS, Mark I, Chen C, Cook AR, Lee HC, Lee VJ, Lin RTP, Tambyah PA, Goh LG (2010) Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. PloS one 5(4):e10,036

Papaspiliopoulos O, Roberts GO, Sköld M (2007) A general framework for the parametrization of hierarchical models. Statistical Science pp 59–73

Parry M, Gibson GJ, Parnell S, Gottwald TR, Irey MS, Gast TC, Gilligan CA (2014) Bayesian inference for an emerging arboreal epidemic in the presence of control. Proceedings of the National Academy of Sciences 111(17):6258–6262

Pergl J, Müllerová J, Perglová I, Herben T, Pyšek P (2011) The role of long-distance seed dispersal in the local population dynamics of an invasive plant species. Diversity and Distributions 17(4):725–738

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological Economics 52(3):273–288

Plummer M, Best N, Cowles K, Vines K (2006) Coda: Convergence diagnosis and output analysis for MCMC. R News 6(1):7–11, URL `http://CRAN.R-project.org/doc/Rnews/`

Pysek P, Cock MJ, Nentwig W, Ravn HP, et al (2007) Ecology and management of giant hogweed (Heracleum mantegazzianum). CABI

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC (2008) The genomic and epidemiological dynamics of human influenza a virus. Nature 453(7195):615–619

Rannala B (1997) Gene genealogy in a population of variable size. Heredity 78(4):417–423

Roberts GO (1996) Markov chain concepts related to sampling algorithms. In: Markov chain Monte Carlo in practice, Springer, pp 45–57

Robinson F (2007) County Durham and Darlington: where are we now? `http://community.dur.ac.uk/chads/prg/Co%20Durham%20Foundation%20Where%20now%20report.pdf`, accessed: 2014-07-02

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics pp 1151–1172

Salemi M, Vandamme AM (2003) The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press

Sampson C, de Waal L, Child L, Wade P, Brock J, et al (1994) Cost and impact of current control methods used against heracleum mantegazzianum (giant hogweed) and the case for instigating a biological control programme. Ecology and Management of Invasive Riverside Plants pp 55–65

Sellke T (1983) On the asymptotic distribution of the size of a stochastic epidemic. Journal of Applied Probability pp 390–394

Shapiro B, Ho SY, Drummond AJ, Suchard MA, Pybus OG, Rambaut A (2011) A Bayesian phylogenetic method to estimate unknown sequence ages. Molecular Biology and Evolution 28(2):879–887

Shaw MW (1995) Simulation of population expansion and spatial pattern when individual dispersal distributions do not decline exponentially with distance. Proceedings of the Royal Society of London Series B: Biological Sciences 259(1356):243–248

Skvortsov A, Ristic B (2012) Monitoring and prediction of an epidemic outbreak using syndromic observations. Mathematical biosciences 240(1):12–19

Spiegelhalter D, Best N, Carlin B, Van Der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B 64(4):583–639

Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by Yule-type speciation models. Mathematical biosciences 170(1):91–112

Ster IC, Singh BK, Ferguson NM (2009) Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. Epidemics 1(1):21–34

Stone M (1997) Discussion of Aitkin (1997). Statistics and Computing 7:263–264

Streftaris G, Gibson GJ (2004a) Bayesian analysis of experimental epidemics of foot-

and-mouth disease. Proceedings of the Royal Society of London-B 271(1544):1111–1118

Streftaris G, Gibson GJ (2004b) Bayesian inference for stochastic epidemics in closed populations. Statistical Modelling 4(1):63–75

Streftaris G, Gibson GJ (2012) Non-exponential tolerance to infection in epidemic systems-modeling, inference, and assessment. Biostatistics 13(4):580–593

Tanaka MM, Francis AR, Luciani F, Sisson S (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173(3):1511–1520

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from dna sequence data. Genetics 145(2):505–518

Turner PC (2005) Molecular biology. Garland Science

Valleron AJ, Boelle PY, Will R, Cesbron JY (2001) Estimation of epidemic size and incubation time based on age characteristics of vCJD in the United Kingdom. Science 294(5547):1726–1728

Vilà M, Basnou C, Pyšek P, Josefsson M, Genovesi P, Gollasch S, Nentwig W, Olenin S, Roques A, Roy D, et al (2009) How well do we understand the impacts of alien species on ecosystem services? A pan-european, cross-taxa assessment. Frontiers in Ecology and the Environment 8(3):135–144

Wald A (1973) Sequential analysis. Courier Corporation

Walters RJ, Hassall M, Telfer MG, Hewitt GM, Palutikof JP (2006) Modelling dispersal of a temperate insect in a changing climate. Proceedings of the Royal Society B: Biological Sciences 273(1597):2017–2023

Watson RK (1972) On an epidemic in a stratified population. Journal of Applied Probability pp 659–666

Yang Z (1994) Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution 39(1):105–111

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution 11(9):367–372

Yang Z (2006) Computational molecular evolution, vol 284. Oxford University Press Oxford

Ypma R, Bataille A, Stegeman A, Koch G, Wallinga J, Van Ballegooijen W (2012)

Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings of the Royal Society B: Biological Sciences 279(1728):444–450

Ypma R, Van Ballegooijen W, Wallinga J (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 113

Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philosophical Transactions of the Royal Society of London Series B, Containing Papers of a Biological Character 213:21–87

Zhang L, Diaz RS, Ho DD, Mosley JW, Busch MP, Mayer A (1997) Host-specific driving force in human immunodeficiency virus type 1 evolution in vivo. Journal of Virology 71(3):2555–2561